NBER WORKING PAPER SERIES

ESTIMATING THE VALUE OF OFFSITE TRACKING DATA TO ADVERTISERS: EVIDENCE FROM META

Nils Wernerfelt
Anna Tuchman
Bradley Shapiro
Robert Moakler

Working Paper 32765
http://www.nber.org/papers/w32765

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
August 2024

Wernerfelt and Moakler were employees of Meta when this research was conducted and the latter owns stock in the company. Meta was able to review this publication for proprietary, trade-secret, or non-aggregated information that could potentially identify any individual(s), but did not have the right to restrict publication based on the results or content of the findings. Neither Tuchman nor Shapiro have material interest in or financial relationship with Meta. Mr. Shapiro's research is supported by the True North Endowment Fund at the University of Chicago Booth School of Business. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Estimating the Value of Offsite Tracking Data to Advertisers: Evidence from Meta
Nils Wernerfelt, Anna Tuchman, Bradley Shapiro, and Robert Moakler
NBER Working Paper No. 32765
August 2024
JEL No. L15,L40,L49,L59,M30,M31,M37,M38

ABSTRACT

Third-party cookies and related 'offsite' tracking technologies are frequently used to share user data across applications in support of ad delivery. These data are viewed as highly valuable for online advertisers, but their usage faces increasing headwinds. In this paper, we quantify the benefit to advertisers from using such offsite tracking data in their ad delivery. With this goal in mind, we conduct a large-scale, randomized experiment that includes more than 70,000 advertisers on Facebook and Instagram. We first estimate advertising effectiveness at baseline across our broad sample. We then estimate the change in effectiveness of the same campaigns were advertisers to lose the ability to optimize ad delivery with offsite data. In each of these cases, we use recently developed deconvolution techniques to flexibly estimate the underlying distribution of effects. We find a median cost per incremental customer at baseline of $38.16 that under the median loss in effectiveness would rise to $49.93, a 31% increase. Further, we find ads targeted using offsite data generate more long-term customers per dollar than those without, and losing offsite data disproportionately hurts small scale advertisers. Taken together, our results suggest that offsite data bring large benefits to a wide range of advertisers.

Nils Wernerfelt
Kellogg School of Management
2211 Campus Drive
Evanston, IL 60208
nils.wernerfelt@kellogg.northwestern.edu

Anna Tuchman
Northwestern University
Kellogg School of Management
2211 Campus Dr
Evanston, IL 60208
anna.tuchman@kellogg.northwestern.edu

Bradley Shapiro
Booth School of Business
University of Chicago
5807 South Woodlawn Avenue
Chicago, IL 60637
and NBER
Bradley.Shapiro@chicagobooth.edu

Robert Moakler
Meta
rmoakler@meta.com

1 Introduction

Digital advertising now constitutes a majority of total advertising spending (Cramer-Flood 2021). From an advertiser's perspective, one of the primary benefits of digital over other channels is the ability to use detailed user data to match ads to consumers.

A challenge in using such data for advertising is that often the data are generated on a different website or mobile application from where the ad delivery will occur. For example, an advertiser may want to place an ad in the Facebook app based on a user's actions in Safari. Nearly every major advertising platform today offers a way for such 'offsite' data to be tracked with third-party cookies and sent back to the platform to be used in ad delivery in real time. Uptake of these tools has been high: in a 2021 survey of marketing professionals in the US, 83% reported using third-party cookies and 51% said such cookies are 'very important' for their current marketing strategy (Innovid 2021).

At the same time, the use of such offsite tracking data faces increasing headwinds from a combination of product and regulatory changes. Apple's roll out of their "Ask App Not to Track" feature in iOS 14.5 made it easier for users to opt out of being tracked by such methods, and similar changes are on the horizon for Android users. Further, policies similar to the General Data Protection Regulation (GDPR) that are being considered in India, Brazil, and other countries may also restrict advertisers' abilities to use offsite data. Such changes can have an enormous impact on the advertising industry: analysts estimate that major advertising platforms lost more than $140 billion in collective valuation after the roll-out of iOS 14.5 (Hackett and Harty 2021).

These changes to the data sharing ecosystem are often motivated by an appeal to consumer privacy. However, there is relatively little evidence on how such changes may affect other important stakeholders. For example, users may value their privacy, but protecting it in this manner may come at a cost to advertisers in the form of reduced advertising effectiveness and to users in the form of seeing less relevant or potentially more ads.

In this paper, we focus on understanding one piece of this puzzle – quantifying the value of offsite data to advertisers. In particular, we estimate how advertising effectiveness would change under a loss of offsite data. To do this, we conduct a large-scale, randomized experiment with more than 70,000 advertisers on Facebook and Instagram. Offsite data can be used by ad platforms in a variety of different ways; we focus on quantifying the value of one primary use case that is offered at Meta and nearly every other major ad platform today. At Meta, our use case is referred to as 'offsite conversion optimization.'[1] Under offsite

[1]Offsite conversion optimization can refer to different types of 'conversion' events that happen offsite to Meta (e.g., website views, email sign ups, donations, etc.). We focus on purchase events in our paper.

conversion optimization, the platform collects offsite data on purchases of the focal product in real time and uses that as training data to predict which users are likely to convert. Meta then dynamically adjusts the target audience, aiming to spend the advertiser's budget on impressions with a high predicted probability of conversion. Advertisers on Meta are required to specify an optimization goal for every campaign they run. Given the vast majority of sales happen offsite from Meta (e.g., on the advertiser's website), offsite conversion optimization is a very popular and logical choice for advertisers interested in driving sales. We discuss this use case in greater detail in the next section.

We design an experiment to estimate two main quantities of interest. First, we estimate the baseline effectiveness of advertising campaigns that use Meta's offsite conversion optimization. To do this, we employ a novel experimentation infrastructure that generates advertising experiments on top of live campaigns. For each campaign in our sample, we randomly hold some users out from seeing its ads. Comparing purchase outcomes across users who were sent versus withheld from seeing the focal ads, we can estimate the cost per incremental customer for each campaign.

Second, we estimate how the effectiveness of each campaign would change if ad delivery could only be optimized for an outcome that Meta observes on its own platform – ad clicks. To estimate this quantity, we take a small fraction of traffic from the same campaigns and optimize its delivery for clicks instead of purchases. We also hold some users out from seeing ads from these modified campaigns, allowing us to estimate how much less effective the same campaigns would be were the advertiser to stop optimizing delivery for purchase outcomes. In short, for each campaign in our sample, we generate two experiments that estimate cost per incremental customer if the campaign were optimized for purchases and if it were optimized for clicks. We repeat this process across all of the 70,000+ campaigns in our sample.

To obtain our main results, we conduct meta-analyses of our experiments. In our data the empirical distribution of treatment effects across experiments is highly skewed, motivating an interest in not only average effects but quantiles as well. Further, given that we experiment on a fraction of traffic from live campaigns, most of our experiments are relatively small and individually under-powered to draw conclusions in isolation. To accommodate these aspects of our data and flexibly estimate the latent distribution of effects, we use the empirical Bayes deconvolution method from Efron (2016). More standard meta-analytic approaches – such as fixed effects or normally distributed random effects – do not fit the data as well.

Overview of Results. Our analysis yields several key results. First, looking at baseline advertising effectiveness, we estimate a median cost per incremental customer of $38.16, with

10th and 90th percentiles of $5.86 and $241.95, respectively.[2] These estimates are ballpark consistent with many customer acquisition cost (CAC) benchmarks, though – reflecting our minimally selected sample and focus on incremental effects – ours tend to be on the higher side.[3]

Second, we re-estimate our main specification on the within-campaign difference in ad effectiveness when the campaign is optimized for clicks instead of purchases. We estimate that under click optimization, the median advertiser earns 6.2 fewer incremental customers per $1,000 of ad spend. This implies that, at the median, optimizing for clicks increases the cost per incremental customer to $49.93 from a median cost of $38.16 when optimizing for purchases. This represents a 31% increase in the cost of acquiring incremental customers. Further, about 90% of the estimated within-campaign differences lie below zero, suggesting that the vast majority of the advertisers in our sample would see a decrease in ad effectiveness if they were unable to run purchase-optimized campaigns.

Last, we conduct two additional analyses: we document heterogeneity in the effects on small versus large scale advertisers and we analyze effects on users' long term purchasing behavior. For the former, we compare results across advertisers who are above versus below median in past ad spend on Meta. Though advertisers with small levels of spend are of interest in their own right, past surveys show that marketing budgets tend to be correlated with measures of firm size (e.g., Moorman (2023)). Further, small businesses are especially policy relevant since they rely more heavily on digital advertising than large businesses do, and thus may be more exposed to any shocks in advertising effectiveness (Kerrigan and Keating 2019, Herhold 2019). In our experiment, we find evidence that small scale advertisers are hurt much more by a loss of offsite data than are large scale advertisers: the level change in ad effectiveness is much greater for the former than it is for the latter. While our data do not allow us to unpack the mechanisms behind this difference, we discuss this descriptive finding in more depth in Section 5.2.3.

In a final analysis, we look at the purchasing behavior of users during a window six months after our experiment ran. We find evidence that purchase-optimized ads generate substantially more long term customers per dollar than click-optimized ads do. We discuss the different potential sources of this result and interpretations of what it means in terms of consumer welfare. On one hand, this evidence is consistent with a state of the world where using offsite data improves the match quality between products and consumers – showing consumers ads for products that are a better fit increases the number of consumers who

[2]Since our estimation procedure allows us to recover the entire latent distribution of effects, these percentiles refer to points across that distribution. They do not constitute a confidence interval.

[3]Cost per incremental customer is similar to CAC but distinct, as cost per incremental customer does not take into account whether the customers are new.

choose to purchase long term and, as a result, finds more valuable customers for advertisers as well. On the other hand, it is possible that these long-term effects are the result of subsequent ad re-targeting or state-dependent demand. Thus, we view these findings as suggestive but non-dispositive evidence around the consumer welfare implications of offsite data.

Our results come with caveats. First, our experiment is intentionally designed to be partial equilibrium. The counterfactual we consider is: if an advertiser who uses purchase data to optimize delivery of their Meta ads unilaterally switches to optimizing for clicks instead, how much less effective will their ads be? This is not the counterfactual of how the world would look if all advertisers simultaneously stopped optimizing campaigns for purchases. In particular, our counterfactual does not capture general equilibrium effects that might come from broader changes, such as R&D into new targeting procedures, ad price changes, or endogenous entry/exit of advertisers, platforms, or users. While these equilibrium outcomes are of high interest, it is challenging to design an experiment that would capture such effects. Instead, we provide clean, experimental identification of partial equilibrium results on the first-order valuations of offsite data to advertisers. Further, our results do not indicate whether optimizing ad delivery for purchases increases social welfare because we do not measure outcomes such as profits to the platform and advertisers, valuations of privacy to consumers, or the welfare benefits to consumers of more relevant ads. Given the scope of the paper, we leave the other pieces of the social welfare question to future research.

Contribution to the Literature. This paper adds to the literature in two main ways.

First, to our knowledge, we provide the largest and most generalizable study of the effects of targeted advertising to date. Our recruited sample contains more than 70,000 advertisers from around the world, more than any other study we are aware of. Further, our sample is minimally selected: historically, only a small fraction of advertisers have run incrementality experiments on the platform using Meta's advertiser-facing experimentation offering. By contrast, our experimental design allows us to experiment across existing live campaigns, generating minimal selection in our sample. Finally, past advertising studies have often not been able to discern whether advertisers in their samples were engaging in direct response or brand advertising, complicating the interpretation of results. We know that the advertisers in our sample all chose to run direct response campaigns, meaning our estimated outcome is thus a relevant metric. Our paper is the first to use this experimentation infrastructure at Meta at this scale, and to our knowledge no other comparable experimentation platform has been used in the advertising literature.

In this light, we build on the marketing literature on measuring advertising effectiveness.

While the existing literature spans decades, it has often by necessity been constrained by issues related to the endogeneity of ad exposure, insufficient statistical power, data quality, and external validity. For example, Gordon et al. (2019) report how non-experimental approaches to measuring ad effectiveness may generate substantial biases, Lewis and Rao (2015) demonstrate how severe power problems can be in this setting, and Kalwani and Silk (1982) highlight subtleties that can arise when using intent to purchase as a proxy for actual purchases. Our paper is able to overcome these challenges with our randomized experimental design, large sample size, and by measuring actual purchasing behavior.[4]

Meta-analyses can overcome some of these issues, for example by pooling information across studies to increase power or obtaining a more representative sample for external validity. Along those lines, three closely related papers are Gordon et al. (2022), Johnson et al. (2017a), and Athey et al. (2023), each of which analyzes hundreds of online advertising experiments (by advertisers who self select into experimentation) to compare measurement from experimental versus non-experimental methods, carryover, and other effects of online display ads, respectively. In contrast, our experiment contains the near universe of advertisers in the target population and not just those that self-select into using measurement tools. As a result, our findings are quite generalizable. Two other large scale, across advertiser studies similar in spirit to ours are Lodish et al. (1995) and Shapiro et al. (2021). However, these studies focus on television advertising and the samples cover hundreds of consumer packaged goods (CPG) advertisers, compared to the tens of thousands of digital advertisers we have in our experiment. Building on this prior work, we view our baseline estimates as making an important contribution to the literature, as they help researchers and advertisers generate an unbiased prior distribution of digital advertising effectiveness.

The second main contribution of this paper is to provide the most comprehensive and generalizable evidence to date on the value of offsite tracking data for ad effectiveness. As mentioned previously, a large share of online advertisers use these data today, but product and regulatory changes may restrict advertisers' ability to do so going forward. At the same time, there is scant large-scale evidence on how advertising effectiveness changes with versus without these data. In designing optimal policy in this space, decision-makers may wish to weigh the effects on users, platforms, and advertisers. We articulate how our findings could help inform policy making in the Discussion section.

This contribution builds on two separate literatures. The first examines the value of

[4] A caveat to our analysis is we are only able to observe purchases that are recorded in Meta's offsite tracking data. Thus, to the extent that online ad exposures drive offline sales or users have opted out of tracking (e.g., after Apple's iOS 14.5 update), our data would lead us to underestimate the effectiveness of the campaigns in our sample. This limitation is not unique to our experiment, and is shared with much of the literature (e.g., Gordon et al. (2019), Tadelis et al. (2023)).

data to different companies. For example, Rossi et al. (1996) find a high value of purchase data for advertisers in the context of a targeted couponing problem, and Dubé and Misra (2023) quantify how much profits increase in their setting when firms can third-degree price discriminate using more user-level covariates.[5] Three other recent online studies run field experiments to analyze the effects of restricting data on firm performance (Sun et al. 2023, Korganbekova and Zuber 2023, Lei et al. 2023). All three find economically meaningful effects in their respective settings, though the data and contexts they focus on (e-commerce product recommendations and clicks on search engine suggestions) are distinct from what we consider.

The second literature is that which examines the impact of privacy regulation on firm-related outcomes. Early work includes Goldfarb and Tucker (2011) who analyze the effects of EU privacy regulation on ad effectiveness and Johnson (2013) who estimates the effects of different privacy policies on advertising prices on an ad exchange. More recently, Aloben-das et al. (2021) analyze the effects of losing third-party cookies on advertiser surplus using non-experimental data and a structural model. There is also a growing literature on the effects of the GDPR, including on investment in technology firms (Jia et al. 2018), firm web traffic and revenue (Goldberg et al. 2021), and mobile app entry and exit (Janssen et al. 2022). An upside of several of these papers compared to ours is that they analyze actual policy interventions. We see our ex-ante analysis of the value of policy-relevant data as complementary to these ex-post evaluations.

Paper Outline. The organization of this paper is as follows: Section 2 describes more background on digital advertising and the relevant context of our paper; Section 3 walks through the experimental design; Section 4 describes our sample; Section 5 presents the main results; Section 6 discusses some broader implications of our findings; and Section 7 concludes.

2 Background

We start by providing some context on digital advertising, including how advertisers deliver ads and measure outcomes. To deliver ads, digital advertising relies on a large set of signals and data, as well as complex modeling processes.

One way to categorize many of the signals and data used in online ad delivery is to bucket

[5] Dubé and Misra (2023) also find that a majority of consumers benefit from more personalized pricing in their study. This finding is similar in spirit to the interpretation of our long term results as reflecting a consumer side benefit to ads delivered with offsite data, though we acknowledge that other interpretations are possible and we cannot draw definitive conclusions regarding consumer welfare from our experiment.

them into first-party or third-party data. The specific data that falls into each category will vary across companies in the ads ecosystem. For example, at Meta, data collected on the platform such as a user's time spent on Instagram and the number of comments they write on Facebook are considered first-party data for Meta. These data are generated on the platform and thus Meta has direct access to the information.

On the other hand, data such as websites visited in Safari or purchases made in a retailer's mobile application are all generated outside of Meta's own applications and would be considered third-party data for Meta. Meta primarily obtains this data to inform ad delivery via tools such as the 'Meta pixel' and 'Meta app software development kit' (SDK).[6]

Pixels are a few lines of JavaScript code or actual 1-by-1 pixel images that can be placed on websites and programmed to 'fire' when individuals take certain actions (e.g., purchase a product or view a page). App SDKs are programming libraries that provide similar functionality in mobile applications. Consider the case of a user who visits a shoe retailer's website: if the retailer has installed a Meta pixel on their website and the user buys a product, the pixel will fire and tell Meta about the purchase event. Similarly, if the user had instead used the retailer's mobile application to complete their purchase, Meta could log this sale if the retailer had implemented the Meta app SDK in their app. Under the hood, pixels and SDKs work in conjunction with third-party cookies and related technologies to transmit data back to the platform. Hence, without functionality like that provided by third-party cookies, the data sharing as described above breaks down.

Pixels and the app SDK are the main source of offsite data advertisers use to optimize the delivery of their ads on the Meta platform. We include data from both sources in our analysis, though for the remainder of this paper we abuse terminology slightly and use 'pixel data' as a shorthand way of collectively referencing offsite data from both. Finally, we note that advertisers must install these tools themselves, so the collection of offsite data that Meta acquires from pixels and SDKs is dependent on advertisers. Advertisers do face an incentive to utilize these techniques, as they can aid in both the optimization and measurement of advertising campaigns as we will discuss next.

Our focus in this paper is on understanding how valuable offsite data are for advertisers on Meta. There are many ways advertisers can use pixel data in their advertising. We do not aim to articulate every use case and measure the corresponding effects; rather, we focus on arguably the main way advertisers use such data today and estimate what would happen if they were forced to abandon that practice. We study the use of pixel data to

[6]Pixels and app SDKs were created by advertising platforms to improve advertising on those platforms. Meta, Snap, TikTok, and other large advertising platforms all have their own pixels and SDKs. The Meta tools only send data to Meta and not to other companies.

perform 'offsite conversion optimization,' and we compare this practice to a counterfactual in which advertisers must instead perform 'onsite click optimization.' We describe each of these practices in more detail next.

2.1 Offsite Conversion Optimization

Advertisers have to specify an objective when they set up ad campaigns on Meta. Why does Meta require this for all campaigns? Consider the following example. Suppose an advertiser wants to generate more likes on their Facebook Page. Given a target audience, creatives, and a budget, Meta could distribute the ads uniformly across users in the target audience until the budget runs out.

However, note that over the course of the campaign, Meta will observe which users actually like the Page. Suppose that in a given target audience, only women like the Page. After observing this, it would likely be inefficient for Meta to keep distributing the ads uniformly amongst the target population because part of the advertiser's budget would be spent on impressions to men, who are unlikely to take the desired action. Further, such ad delivery would presumably not be best for users, since men would be exposed to ads with which they are not interested in engaging. Instead, Meta could improve upon uniform delivery by training a model that predicts the probability a user would like the Page in question and using that model to dynamically inform which users should see the ads. In this case, such a model would lean towards delivering ads to women, thereby reducing inefficiencies for both the advertiser and users. This process of refining ad delivery within the campaign's target audience according to the advertiser's stated objective is called delivery 'optimization' at Meta.[7]

In the above example, Page likes are a first-party outcome for Meta: this behavior occurs within the platform. For many advertisers, however, the actions they care about happen offsite from Meta. For example, purchases rarely occur on Meta's platform – they are usually made in web browsers or other mobile applications. This is where pixels play an important role. If an advertiser installs a pixel, it can transmit data to Meta on offsite behavior. Meta can then use that data as the left-hand side variable in the prediction problem, dynamically finding users to show the ads to who are likely to take the offsite action the advertiser cares about. This process is called 'offsite conversion optimization' and measurement of these offsite outcomes plays a central role in ad delivery.

In this paper, we focus on campaigns that are optimized to drive purchases, which is the

[7]The actual training and deployment of these delivery models is more complex than what is described here due to the engineering scale and speed at which digital ad auctions occur, though the intuition is the same. See https://www.facebook.com/business/help/1000688343301256 for a high level overview.

most common offsite objective that advertisers use. Intuitively, one can understand why this approach is attractive for direct response advertisers. They may not know who their target audience is or have the right features to target them on, but Meta can use a data driven approach to help optimize the matching problem. Indeed, if an advertiser wants to maximize sales, offsite conversion optimization for purchases is the logical choice. However, there is no guarantee that this method maximizes incremental purchases. Meta could just show the ad to people who were already going to purchase the product, so in isolation, it is unclear how many of the ad impressions generate incremental sales. We structure our experiment to address this concern and identify the incremental effects of interest.

Finally, we emphasize that almost every major online ad platform offers a similar way to optimize ads for offsite outcomes. For example, TikTok, Snapchat, Pinterest, Twitter, and others all have variants of pixels that their advertisers can install, and all offer ad delivery optimization for offsite events. As a result, a large share of digital advertising today relies on pixels, and while our experiment is conducted on Meta, we believe our results are informative for the broader industry.

2.2 Click Optimization

If an advertiser with a pixel installed lost the ability to optimize delivery for offsite outcomes, what would they substitute to? We assume they would instead optimize for click-through rate. In particular, "clicking on the ad" would replace "making a purchase" as the objective to optimize when allocating ads.[8]

There are three main reasons why we focus on this counterfactual. First, the lowest outcome in the conversion funnel that Meta observes onsite is the user clicking the ad to be taken offsite. Hence, clicking is as 'close' to purchase as one can get in the funnel when relying on Meta first-party data alone. Second, the two main optimization types used by direct response advertisers on Meta are offsite conversion optimization and click optimization, and one of the most common questions advertisers ask Meta's salespeople is which one of these two to use, further suggesting this is the relevant choice set. Finally, other advertising platforms also offer click-based delivery optimization, so the counterfactual we consider is also relevant outside of Meta.

[8]This is also known as 'link click optimization' at Meta, referring to optimizing for users who click the link on the ad to leave Meta.

3 Experimental Design

Context. Our objective is to estimate (i) the effectiveness of offsite purchase-optimized ads and (ii) how that effectiveness would change if ad delivery were optimized for onsite clicks instead. Each of these outcomes is hard to estimate with observational data. We highlight two main reasons why below.

First, when estimating online advertising effectiveness, past work has shown that non-experimental techniques can lead to substantial bias(Gordon et al. 2022). The delivery optimization procedure dynamically and non-randomly shows ads to individuals who are likely to take an advertiser-specified action, and this process induces selection in advertising exposure that is empirically hard to correct for ex post. Hence, we leverage experimental variation to isolate the causal advertising effects of interest.

Second, even if we cleanly estimated ad effectiveness for existing purchase-optimized and click-optimized campaigns, we could not simply compare the two to estimate the change in effectiveness attributable to offsite data. This is because the advertisers, creatives, and target audiences differ across campaigns that use the two optimization procedures. For example, advertisers who optimize delivery for purchases may use more compelling creatives or operate in different product categories than advertisers who choose to optimize for clicks, meaning we could not disentangle whether any difference in ad performance were due to offsite data or heterogeneity on other dimensions. We address these challenges by designing and implementing a study that generates experimental estimates of advertising effectiveness under both purchase and click optimization for each campaign in our sample.

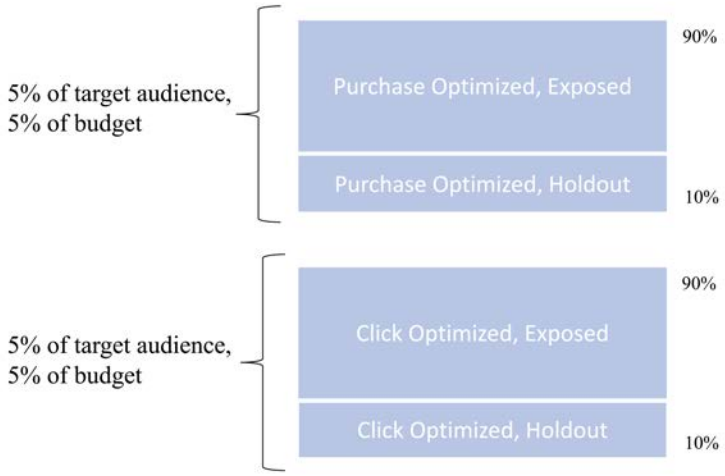
Implementation. Figure 1 outlines our experimental design for a single campaign. To generate our final sample, we repeat this procedure for more than 70,000 campaigns in parallel.[9] We describe each step of the experiment below.

Consider an ad campaign whose objective is to drive sales as logged by the advertiser's pixel. First, we create experimental variation to estimate the effectiveness of this purchase-optimized campaign. We randomly take 5% of the campaign's target audience and allocate 5% of the overall ad budget to be spent on those users. We refer to this 5% sample as the Purchase Optimized group. We randomly assign 10% of the users in the Purchase Optimized group to a 'holdout' group that is ineligible to see the focal ads, and the other 90% of users form the 'exposed' group.[10] Within the Purchase Optimized group, Meta dynamically

[9] In Web Appendix A we describe how we define a campaign, and in Web Appendix E we walk through the experiment for three example campaigns from our sample.

[10] We create the holdout and exposed groups by leveraging the same infrastructure used in Meta's advertiser-facing 'lift' tool (Gordon et al. 2019).

Figure 1: Overview of the core experimental design.



determines which users to show the ads to leveraging the users' predicted probability of purchasing the focal product. Ad impressions on Meta are allocated using an auction, so the advertiser must bid on impressions for these targeted users. If an ad from the focal campaign wins the auction for a user in the exposed group, Meta sends the ad to the user. If, on the other hand, the user is in the holdout group, Meta withholds the ad and instead sends the ad this user would have received had the focal campaign not existed. We compare outcomes for users who were sent the ad versus those for whom the ad was withheld and the second place ad was sent instead, similar to the approach described in Johnson et al. (2017b).[11] This procedure generates a baseline estimate of ad effectiveness under purchase optimization, where our main outcome of interest is cost per incremental customer.[12]

We also estimate how effective the same ads would be at generating incremental customers under click optimization. To do this, we take another 5% of the target audience and budget and allocate them to a Click Optimized group. For this segment of users and budget, we deliver the ads in a way that approximates click optimization — the ads are delivered to people who are predicted to be likely to click on the ads. We include an analogous 10% holdout group within the Click Optimized segment of users, allowing us to estimate the incremental effect of the click optimized version of the ads.[13]

[11] The analysis is technically intent-to-treat because even though an ad won the auction, a user may not actually see it for a variety of reasons (e.g., scrolling too fast by it, not scrolling far enough down, or loss of network connection). Hence, we are comparing users who were assigned to treatment and control. We note that the standard output advertisers get from Meta's measurement tools are also intent-to-treat estimates for the same reason.

[12] See Web Appendix C for a derivation of the experiment-level treatment effects and standard errors of cost per incremental customer from the individual-level data.

[13] We need a separate holdout group for each of the purchase and click optimized arms because the optimization

A critical feature of our experiment is that we observe offsite purchase data for all users, including those randomly assigned to the click-optimized exposed arm and the holdout groups. This allows us to conduct a comparison of the same outcome (cost per incremental customer) across arms. One important caveat is that we do not observe purchases made at brick and mortar stores or purchase data for users who had opted out of tracking (e.g., from Apple's ATT prompt).[14,15]

Additional points. We close out this section with a discussion of two additional implementation details.

First, for engineering reasons, it was infeasible to implement Meta's production version of click optimization; we thus relied on a model that uses simpler inputs to predict the relevant click outcomes. To the extent that this model does not perform as well as the production model, our click optimization counterfactual estimate would be biased downward. To investigate this, we ran a separate experiment where we took campaigns that were delivered according to the production click optimization model and switched a fraction of their traffic to be delivered according to the model in our experiment. In practice, the difference in model performance was small: the cost per purchase under the experiment's simpler arm was 10% higher, or about 0.04 standard deviations of our estimated distribution.[16] Hence, we believe the size of any model-induced error is likely small.

Second, we emphasize that the counterfactual we analyze is: if an advertiser who is using offsite conversion optimization switches to click optimization, how much less effective will their ads be? This is distinct from the counterfactual where all pixel data is removed from their ad delivery. The right hand side of the prediction problem has a vast number of inputs, some of which are functions of pixel data, and removing all traces of pixel data from the right-hand side was infeasible from a practical perspective.[17] However, including pixel data and its derivatives as features on the right-hand side should if anything only improve

algorithms endogenously select different target populations.

[14]This limitation is not unique to our experiment, and is shared with much of the literature (e.g., Gordon et al. (2019), Tadelis et al. (2023)). In our results, if we assume this missing data problem affects purchase- and click-optimized campaigns in the same way, our primary statistic of the median effectiveness evaluated at the median loss remains unchanged. To the extent that it may differentially affect click or purchase optimized ads, our estimates would rescale accordingly.

[15]An additional caveat is that while advertisers have stated an objective of sales for all the campaigns in our sample, they may vary in other, unobserved dimensions. For example, some may also care about brand awareness, even if they did not explicitly tell Meta.

[16]To assess the sensitivity of our results to this difference, we re-ran our estimation procedure with the costs in the click arm scaled up by 10%, and found that our main statistic of interest was similar.

[17]For example, models that predict user characteristics could have been calibrated with pixel data, or on-platform activity that feeds into the algorithm could have been influenced by pixel data (e.g., a user sees another ad targeted with pixel data and then behaves differently on platform).

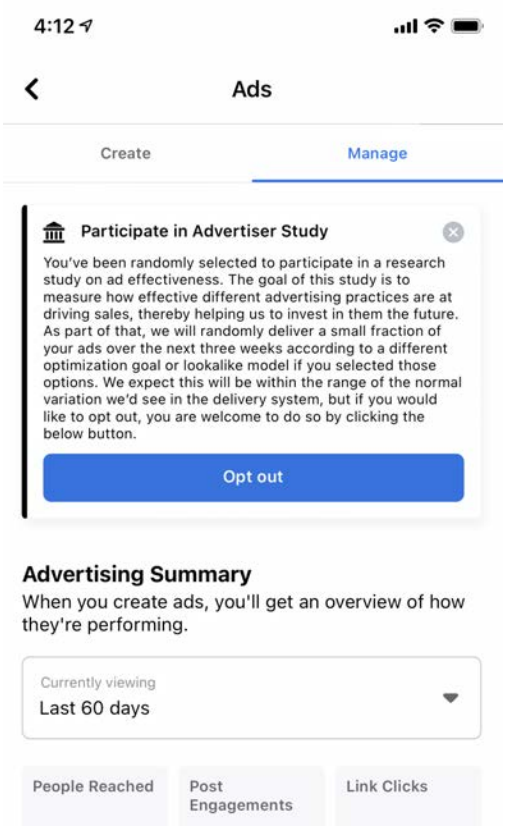
the performance of the click optimization algorithm. Thus, if one were interested in the counterfactual of removing all pixel data from the ad delivery process, our estimate can arguably be thought of as a lower bound.

4 Sample

4.1 Recruitment and Selection

Our study ran for one week in the Fall of 2021 with an aim of recruiting as large and representative a sample as possible. To that end, Meta sent advertisers a recruitment notice that described the experiment and provided them the opportunity to opt out. The notice appeared as a banner in the main surfaces through which advertisers purchase ads on Meta and was translated into the 31 languages most frequently used by the company's advertisers. Figure 2 shows an example of the notice. The notice would appear each time an advertiser logged in (up to three times) and would then disappear.

Figure 2: Example of the recruitment notice sent to advertisers.



All accounts that ran an offsite conversion optimized campaign on Facebook or Insta-

gram during the preceding three months were eligible to receive the notice. This amounted to around four million advertiser accounts. In addition, an account was eligible for our experiment only if they logged in, were exposed to the notice, and chose to not opt out.[18] Among the accounts that were eligible to see the notice, only a subset visited the recruitment interfaces while it was live. Conditional on seeing the notice, 94% of accounts did not opt out. These advertiser accounts ran 187,922 purchase-optimized campaigns during the week of our experiment.

Within that sample, we focus on a subset of the campaigns for our main analysis. Meta recommends that advertisers should not use offsite conversion optimization if their campaign is expected to achieve less than fifty conversions per week. Meta makes this recommendation because the optimization model may not perform well if the left hand side variable is very sparse. A large share of the campaigns in our sample do not hit this minimum, meaning their ad delivery may not be fully optimized. This generates two research questions of interest: how effective is offsite conversion optimization for advertisers who use it, and how effective is it for advertisers who use it as recommended? We focus on the latter question in the main text because we believe it speaks to the potential upside of offsite data, which we see as more relevant to the policy debate and advertisers' marketing decisions. In Web Appendix B, we re-run our main analysis on the full sample.[19]

We cannot directly filter to campaigns that hit the minimum because we only observe conversion events from our experiment, which comprises a fraction of overall campaign traffic. Instead, we restrict to campaigns that met the threshold in expectation: i.e., campaigns that recorded more than $50 \times 0.05 = 2.5$ conversions in the purchase optimized arm, since this arm comprised 5% of total traffic. Conditioning on this minimum reduces the final sample to 70,909 experiments.[20]

[18] Some interfaces through which advertisers purchase ads are not amenable to notices (e.g., command line APIs). Since our notice could not be sent to these surfaces, advertisers who only use such interfaces were not included in our experiment. Only a small fraction of all advertisers use such surfaces, and, as we show, the characteristics of advertisers in our final sample do match well with the overall population.

[19] As expected, cost per incremental customer rises, though the core result on the percent reduction in effectiveness from our counterfactual is of a comparable magnitude.

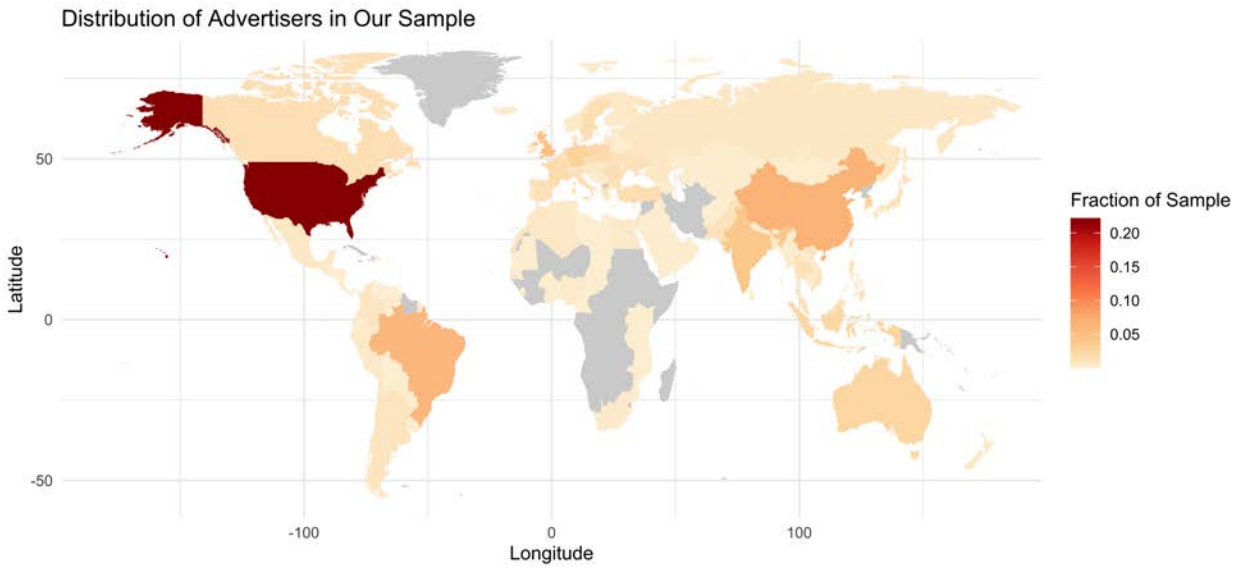
[20] Note that 63% of experiments failed to hit the recommended minimum in expectation. There are several possible explanations for why such a large share do not meet the minimum. For example, advertisers may not know about this recommended minimum, they may believe even sub-optimal purchase optimization is better than any alternative, or they may be overconfident in their purchase volume.

4.2 Characteristics of Advertisers

In this section, we describe the characteristics of the advertiser accounts associated with the 70,909 experiments in our final sample.

Regions. Our final sample includes advertiser accounts based in more than 160 different countries (see Figure 3). The US is the most represented country, with around 22% of the sample, and other major countries include China (7%), Brazil (6%), and India (4%).[21]

Figure 3: Geographic distribution of advertisers in our sample.

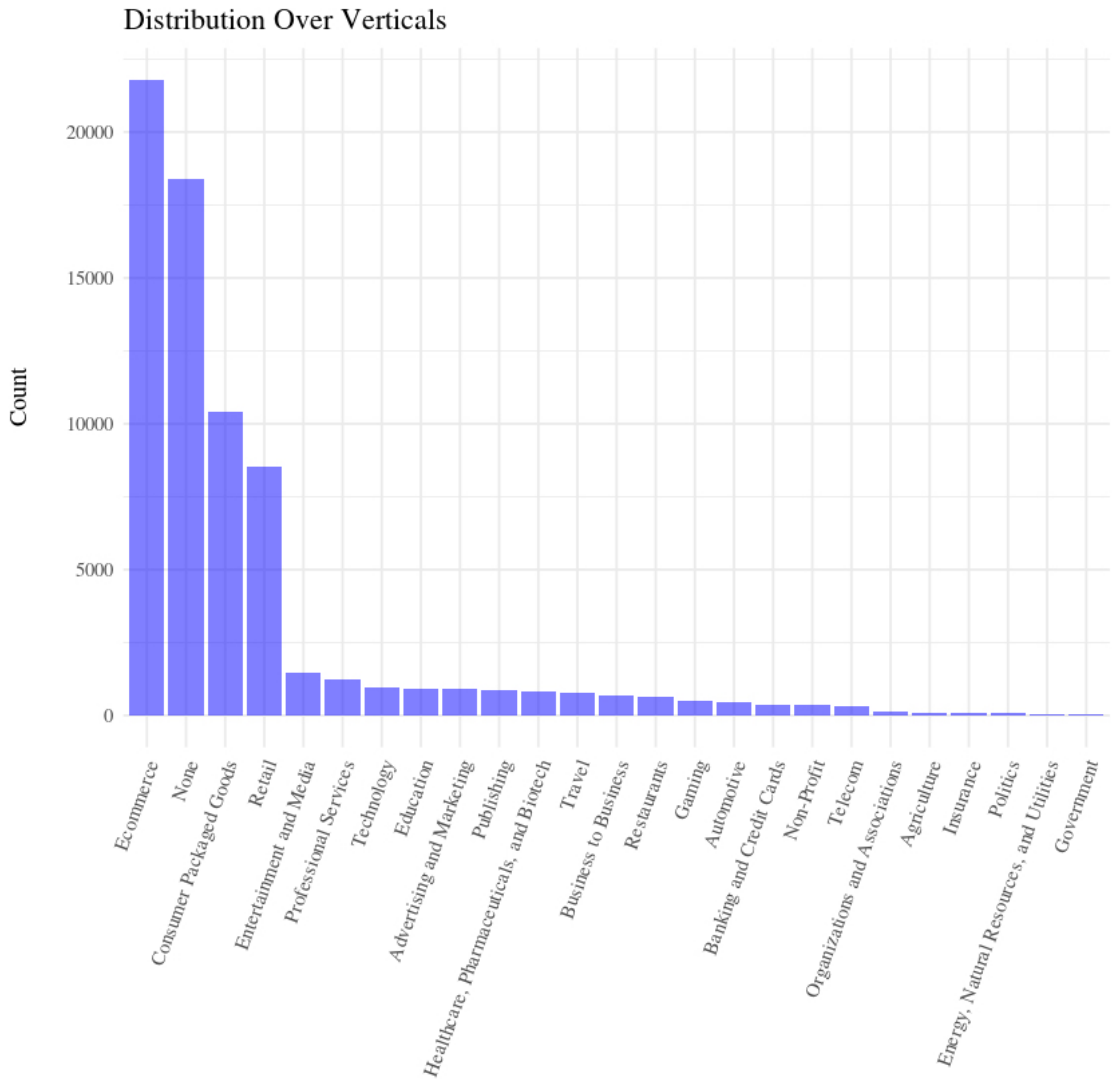


Notes: Grey denotes a country with no advertisers in our sample.

[21] Country refers to the location of the advertiser. Despite the fact that Facebook and Instagram are banned for users in China, there are many Chinese-based advertisers who advertise to users in other countries.

Verticals. Our sample also covers a wide spectrum of industries. The most heavily represented verticals are E-commerce (44%), Retail (20%), and CPG (12%), with a long tail after that (see Figure 4). E-commerce is a broad category that largely consists of the sub-verticals Apparel and Accessories and Durable Household Goods.

Figure 4: Distribution of verticals in the sample.



Notes: Meta uses an internal machine learning model to predict each advertiser account's vertical. 'None' reflects accounts that could not be classified reliably.

Small and Large Scale. We later compare results for 'small' versus 'large' scale advertisers. In practice we implement this by doing a median split on ad spend in the eight weeks preceding our experiment. To provide a sense of the distribution of previous ad spend in our sample, we list the corresponding quantiles and the mean in Table 1. Note both how skewed the distribution is and how many advertisers spend relatively small amounts.

Table 1: Distribution of ad spend for our sample from the previous eight weeks.

	10th	25th	50th	75th	90th	Mean
Ad spend last eight weeks	$845.15	$3,340.97	$13,510.80	$54,842.42	$195,906.15	$104,615.23

Other Characteristics and Representativeness. We conclude our descriptives by highlighting more characteristics of our sample and comparing them to that of the relevant population.

Table 2 shows that the median advertiser in our sample tends to be fairly engaged with the platform: they have had their account for more than a year, the vast majority were active within the week of our data pull, and a relatively large share have their account managed by Meta (a service provided to sufficiently engaged and high spending advertisers). There is an upfront cost of installing a pixel and familiarizing oneself with the technology, which we suspect screens out some of the very new, less engaged accounts.

On the representativeness front, though we tried to minimize it, there was the possibility for selection bias along the recruitment funnel. For example, one may be concerned that the advertisers who benefit the most from offsite data would disproportionately choose to opt out, thereby affecting the representativeness of our results. In Table 2, we show the difference between the characteristics of our final sample and the characteristics of the population of Meta advertisers who used offsite conversion optimization for purchases with more than the minimum number of recommended purchase events during the week of our experiment. Given the size of the populations, many of the characteristics are significantly different across the two groups. However, the magnitudes of the differences are small: the average absolute value of the percentage point difference is 1.5% across all the characteristics in Table 2. This makes us feel better about the representativeness of our sample.

Table 2: Sample representativeness.

	Experimental Sample	Population	Difference
Vertical			
CPG	0.1196	0.095	−0.0246***
	(1.2e-03)	(0.0017)	(0.0021)
Retail	0.1982	0.2059	0.0077***
	(1.5e-03)	(0.0023)	(0.0028)
E-commerce	0.4443	0.387	−0.0573***
	(1.9e-03)	(0.0028)	(0.0034)
Region			
Asia Pacific	0.2977	0.277	−0.0207***
	(1.7e-03)	(0.0026)	(0.0031)
Europe, Middle East, Africa	0.3673	0.386	0.0187***
	(1.8e-03)	(0.0028)	(0.0033)
Latin America	0.0954	0.0921	-0.0033
	(1.1e-03)	(0.0017)	(0.002)
North America	0.2389	0.2428	0.0039
	(1.6e-03)	(0.0025)	(0.003)
Months since first ad creation			
Less than 1	0.0664	0.0834	0.0169***
	(9e-04)	(0.0016)	(0.0019)
1-6 months	0.157	0.1578	0.0008
	(1.4e-03)	(0.0021)	(0.0025)
6-12 months	0.1228	0.125	0.0022
	(1.2e-03)	(0.0019)	(0.0023)
Greater than 12 months	0.6377	0.6339	-0.0039
	(1.8e-03)	(0.0028)	(0.0033)
Other			
Meta Managed Account	0.3716	0.4037	0.032***
	(1.8e-03)	(0.0028)	(0.0034)
Active Last 7 days	0.9787	0.9994	0.0207***
	(5e-04)	(0.0001)	(0.0006)

Notes: Population refers to the set of Meta advertisers who used offsite conversion optimization for purchases and met the minimum number of recommended purchase events during the week of our experiment. The mean difference corresponds to the population mean minus the mean from our experimental sample. *** denotes $p < 0.01$ in a two-sided t-test on the difference. Standard errors in parentheses.

4.3 Summary Statistics of Experiments

Our final sample includes 70,909 experiments. As mentioned in our Experimental Design section, we create our experiments by taking a small fraction of traffic from live campaigns. Given the long tail of small spend by advertisers, this means that we have a large number of small experiments in terms of both total number of users per condition and the number of converters. Table 3 and Table 4 report summary statistics on these metrics for our final sample. These small experiment sizes help motivate our meta-analysis, in which we pool the individual experiments to better estimate the overall distribution of effects.

Table 3: User count per experiment.

Arm	Min	1st	25th	50th	75th	99th	Max	Mean	SD
Purchase-Exposed	10	111	2,264	6,911	18,862	244,160	9,295,122	22,352	83,827
Purchase-Holdout	1	12	251	767	2,095	27,162	1,032,782	2,483	9,314
Click-Exposed	9	112	2,643	8,384	23,438	335,523	9,374,392	28,796	106,136
Click-Holdout	1	12	293	931	2,605	37,336	1,043,232	3,199	11,794

Notes: "Purchase-Exposed" reports summary statistics and quantiles on the number of users who were randomly assigned to the exposed condition within the purchase-optimized experimental segment. "Click-Holdout" reports the analogous statistics for users who were randomly assigned to the holdout group (no exposure to the focal ads) within the click through rate optimized experimental segment. The remaining groups are defined analogously.

Table 4: Number of converters per experiment.

Arm	Min	1st	25th	50th	75th	99th	Max	Mean	SD
Purchase-Exposed	3	3	4	7	16	821	409,681	76	1,977
Purchase-Holdout	0	0	0	0	2	84	44,273	7	211
Click-Exposed	0	0	2	5	14	782	407,069	72	1,963
Click-Holdout	0	0	0	0	1	79	44,214	7	210

Notes: "Purchase-Exposed" reports summary statistics and quantiles on the number of users who were randomly assigned to the exposed condition within the purchase-optimized experimental segment. "Click-Holdout" reports the analogous statistics for users who were randomly assigned to the holdout group (no exposure to the focal ads) within the click through rate optimized experimental segment. The remaining groups are defined analogously.

Randomization Check. Finally, as mentioned previously, we randomly assigned users to holdout and exposed conditions using the same advertiser-facing lift infrastructure that several other papers have used and validated (Gordon et al. 2019, 2022, Athey et al. 2023). In Web Appendix D, we provide a randomization check of user demographics for our own study and find no evidence of bias.

5 Main Results

5.1 Empirical Distributions of Treatment Effects

First, to provide a high level overview of our data, we plot the empirical distributions of treatment effects and provide the associated summary statistics. The key outcomes of interest are (i) the number of incremental converters per dollar for the baseline, purchase optimized condition (Figure 5, Table 5) and (ii) the within-campaign change in the number of incremental converters per dollar when changing from purchase to click optimization (Figure 6, Table 6). The former refers to optimization with offsite data and the latter refers to how much less effective the same campaigns would be under a loss of offsite data. The distributions of (i) and (ii) are the starting points for the two meta-analyses that we describe later in this section.

We highlight a few key takeaways from these distributions. First, the baseline ad effectiveness distribution has more mass above zero and the within-campaign change in ad effectiveness distribution has more mass below zero. This is consistent with both advertising 'working' – advertising attracts customers – and the offsite purchase data being valuable – if advertisers do not target with that data, their ads become less effective.[22] At the same time, there is substantial heterogeneity: across the 70,909 experiments, some campaigns are estimated to be much more effective than others, and some campaigns are hurt more by the loss of offsite data.[23]

Second, these empirical distributions suggest that losing offsite data has a large impact on ad effectiveness. The median campaign gains an incremental converter for every $23.63 spent; under the median loss this would increase to $36.69, a 55% increase.[24]

Finally, we note that the empirical distributions are highly skewed. Whereas most meta-analyses focus on estimating the grand mean across many studies, given the skew in our sample, we believe estimating quantiles of the underlying distribution helps convey additional pertinent information. In addition, given the non-standard shapes of these distributions, we want to rely on minimal parametric assumptions while flexibly estimating the latent

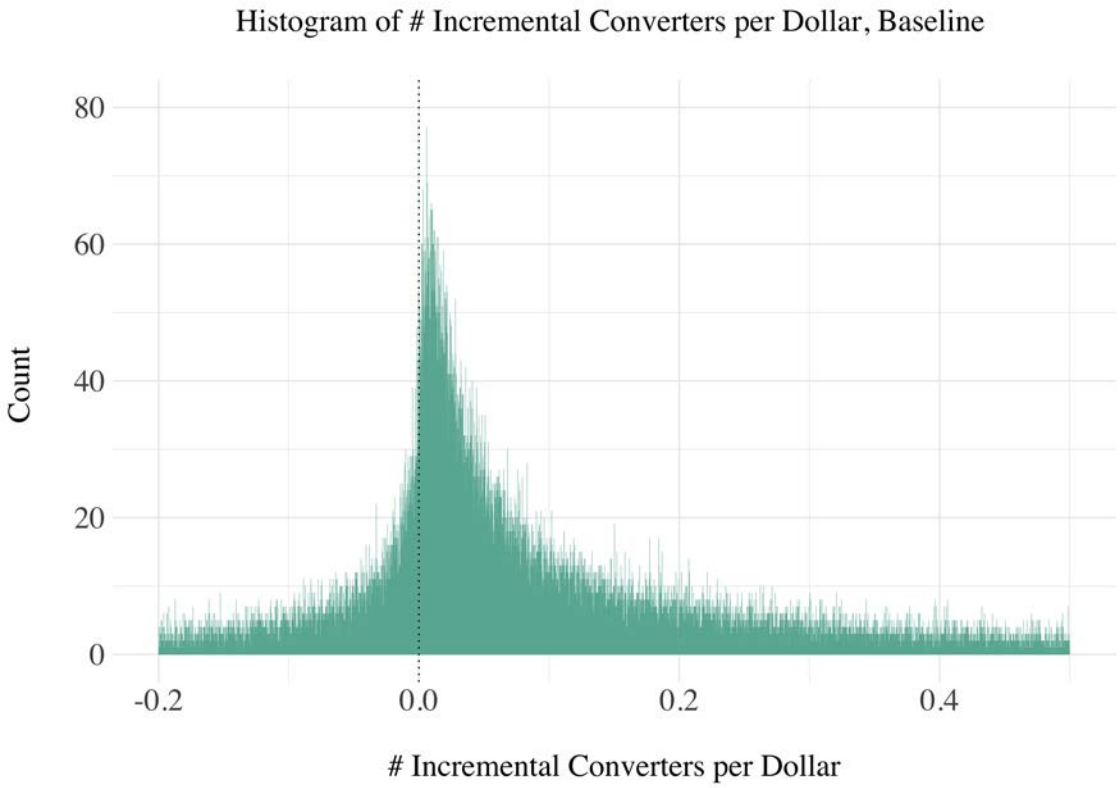
[22] A negative treatment effect for the estimated number of incremental converters per dollar means that each dollar spent on advertising loses customers. This may happen if, for example, consumers have a negative experience with the ad or it leads them to purchase from a competitor, as in Sahni (2016).

[23] Note the massive difference between the extrema and the 1st and 99th percentiles in the tables. As we describe in the Implementation section of our Web Appendix, we drop observations outside the 1st and 99th percentiles in our analyses.

[24] From the table, $23.63 is 1/0.042328 and $36.69 is 1/(0.042328 - 0.01507). More broadly, we report our results in terms of cost per incremental converter because that is more in line with industry terms (e.g., cost per action (CPA)). Further, since many campaigns generated no purchasers, analyzing the data in per converter terms instead of per dollar ad spent would generate many infinite values which are harder to work with.

distributions. These facts motivate our methodology, which we describe next.

Figure 5: Histogram of baseline advertising effectiveness (with offsite data).

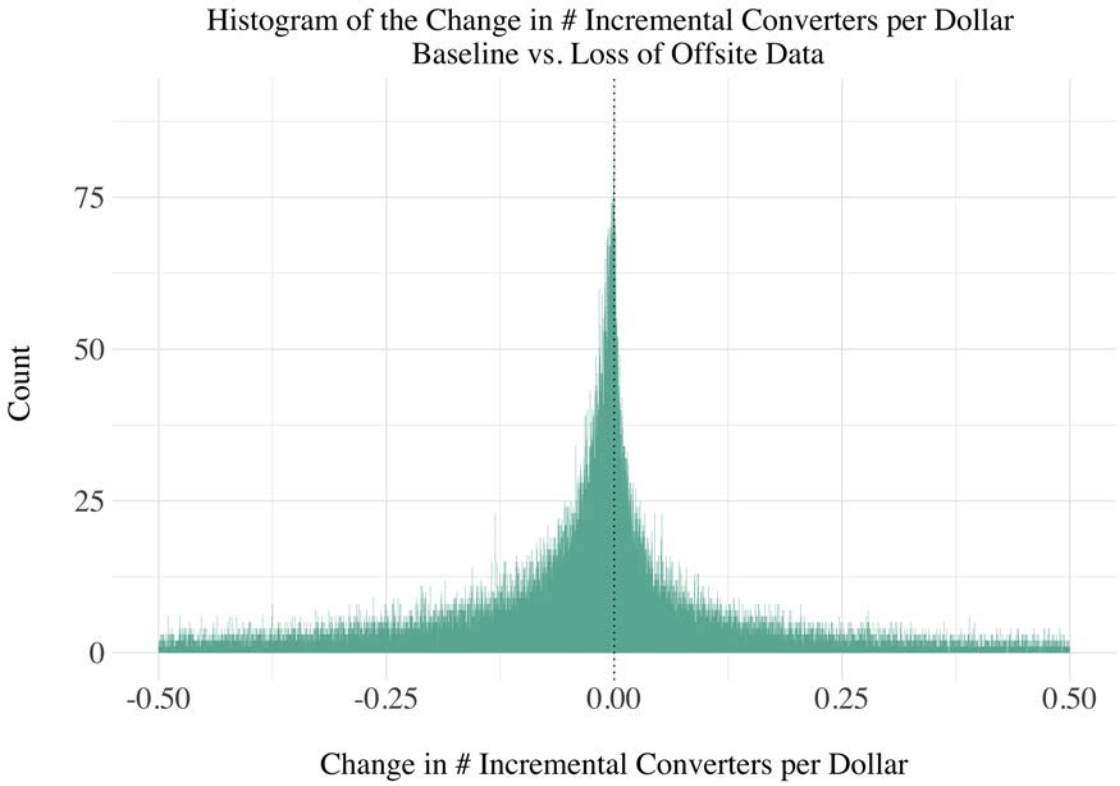


Notes: Bin width is 0.001. Histogram is trimmed at (-0.2, 0.5) for display purposes.

Table 5: Summary statistics and quantiles of the empirical distribution of baseline ad effectiveness.

	Min	1st	25th	50th	75th	99th	Max	Mean	SD
# Incremental Converters per Dollar	-206.53	-3.51	-0.003	0.042	0.199	5.88	5,044.9	0.290	19.42
Cost per Incremental Converter	-$0.005	-$0.285	-$333.3	$23.81	$5.03	$0.17	$0.0002	$3.45	$0.05

Figure 6: Histogram of the within-campaign change in advertising effectiveness.



Notes: Difference computed as number of incremental converters per dollar under click optimization minus the same outcome under purchase optimization. Bin width is 0.001. Histogram is trimmed at (-0.5, 0.5) for display purposes.

Table 6: Summary statistics and quantiles of the empirical distribution of the within-campaign change in ad effectiveness.

	Min	1st	25th	50th	75th	99th	Max	Mean	SD
# Fewer Incremental Converters per Dollar	-1,516	-6.21	-0.133	-0.015	0.039	5.25	13,257	0.066	50.33

Notes: Difference computed as number of incremental converters per dollar under click optimization minus the same outcome under purchase optimization.

5.2 Meta-Analysis

5.2.1 Estimation Procedure

We estimate the latent distribution of effects using empirical Bayes methods (Efron 2014, 2016). We provide a detailed description of the methodology and implementation details in

Web Appendix A and focus on the high level intuition below. We discuss the meta-analysis of baseline ad effectiveness, but the intuition holds equally for the meta-analysis of the change in effectiveness.

Each experiment in our sample provides a noisy estimate of the true, unobserved treatment effect generated by that specific ad campaign. The campaigns in our sample may also differ in their true effectiveness, so across all the experiments in our sample, there is a distribution of these true latent effects. In standard meta-analytic terms, the latter is called the 'higher' level distribution and the former are the 'lower' level distributions. Our goal is to estimate the higher level distribution using the empirical treatment effects and standard errors as inputs with as few parametric assumptions as possible.

We estimate a flexible functional form for the higher level distribution in a computationally tractable way using empirical Bayes (Efron 2014, 2016). For the lower level distributions, we make a normality assumption. This is simply an appeal to the Central Limit Theorem – we impose the assumption that our experiments are sufficiently large that the treatment effects are normally distributed around the true, unobserved value and our variance estimates have converged. We assess the robustness of our results to this assumption in Web Appendix A.

For the higher level distribution, our approach relies on the observation that the exponential family of probability distributions is both very flexible and amenable to computational analyses. Further, we can use splines to reduce the estimation procedure from optimizing over distribution space to optimizing over weights on spline bases. Combining these ideas, we parameterize the higher level distribution assuming it belongs to this flexible family and estimate it using maximum likelihood estimation over weights on spline bases.

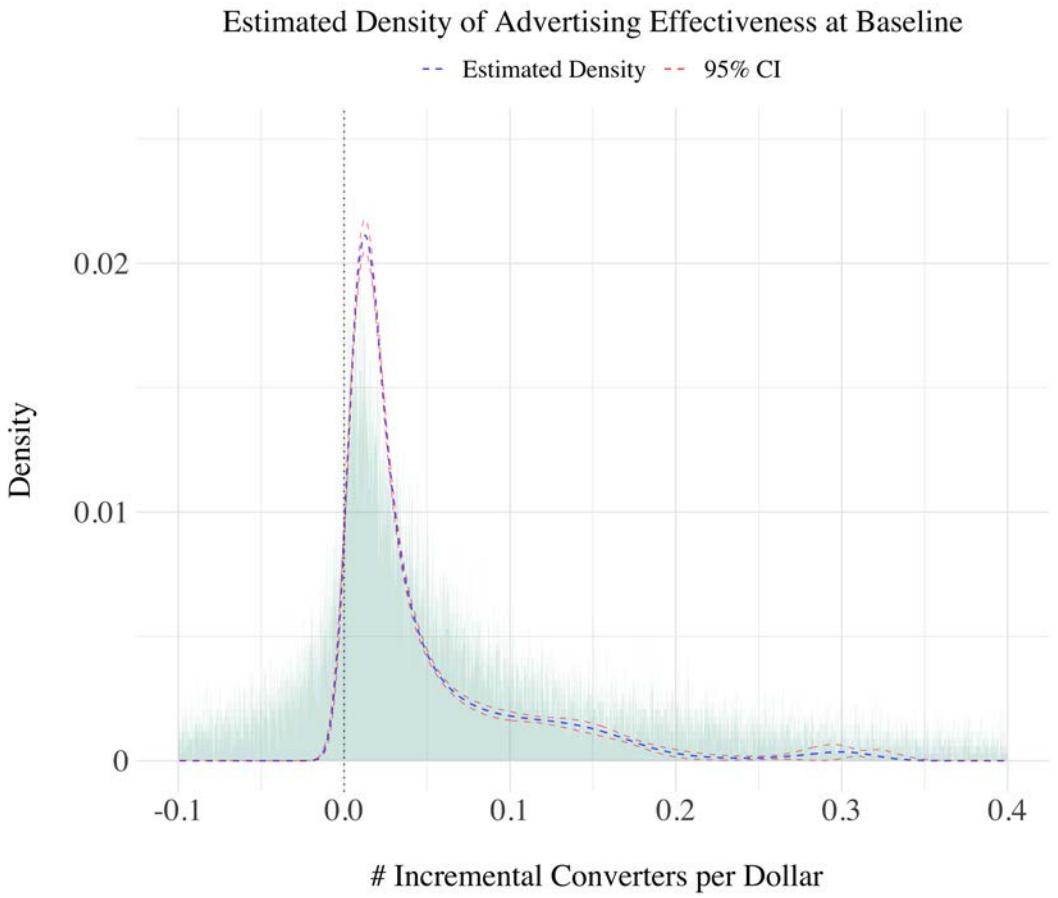
Summarizing the approach at a broad level, we use meta-analytic techniques to recover the latent distribution of effects with minimal parametric assumptions while maintaining computational feasibility.

5.2.2 Overall Results

Baseline Estimates. Our first set of results characterize the estimated distribution of advertising effectiveness for this large sample of advertisers on Meta. In Figure 7 we plot the fitted model overlaid on the histogram of treatment effects, and in Table 7 we provide summary statistics of the fitted distribution.[25]

[25]In Web Appendix F, we explore effect heterogeneity across the main verticals in our sample: E-commerce, Consumer Packaged Goods, and Retail.

Figure 7: Estimated distribution of the number of incremental converters per dollar.



Notes: The blue line represents the estimated distribution of the number of incremental converters per dollar across our sample. The dashed red lines show the 95% confidence intervals calculated from Efron (2016). The empirical histogram of treatment effects is shown in green as a rescaled density with a binwidth of 0.001.

Table 7: Quantiles and mean for the estimated distribution of baseline advertising effectiveness.

	10th	25th	50th	75th	90th	Mean
# Incremental Converters per $1,000	4.1	11.9	26.2	75.8	170.8	105.9
Cost per Incremental Converter	$241.95	$83.70	$38.16	$13.19	$5.86	$9.44

Notes: The top row reports estimates from the distribution in Figure 7 multiplied by $1,000. The second row reports the distribution of the inverse, expressed in terms of dollars per incremental converter.

The empirical histogram of treatment effects in Figure 5 has a sizable mass below zero. From a theoretical standpoint, it would be surprising if a large share of firms were indeed spending money and losing customers. Consistent with that, the fitted density shrinks the negative mass substantially, suggesting that the negative treatment effects are largely noise. In addition, consistent with past literature that has found small effects of advertising (e.g., Lewis and Rao (2015)), we find a large mass just to the right of zero. Finally, we note that, as expected, the empirical Bayes estimator shrinks the overall distribution, as one can see by comparing the quantiles in Table 5 and Table 7.

Table 7 shows that the median campaign in our sample has a cost per incremental converter of $38.16. This is substantially higher than the mean of $9.44, reflecting the skew in the distribution and demonstrating how the results can shift meaningfully when considering quantiles.[26] In a similar vein, had we used a less flexible approach and assumed – as many meta-analyses do – a normal higher level distribution, our median estimate would have been substantially different.[27] We note that our median estimate is higher than many related industry benchmarks, which we interpret as reflecting differences in underlying samples and methodologies.[28]

Finally, Figure 7 shows that our confidence intervals are quite tight around the fitted distribution.[29] We suspect that the bump around 0.3 is due to noise, and the larger standard errors around that region seem consistent with that.

Change in Effectiveness. We now turn to the second main contribution of the paper and look at the estimated change in advertising effectiveness when moving from purchase optimization to click optimization. In Figure 8 we plot the fitted model overlaid on the histogram of estimated effects, and in Table 8 we report summary statistics from the fitted distribution.

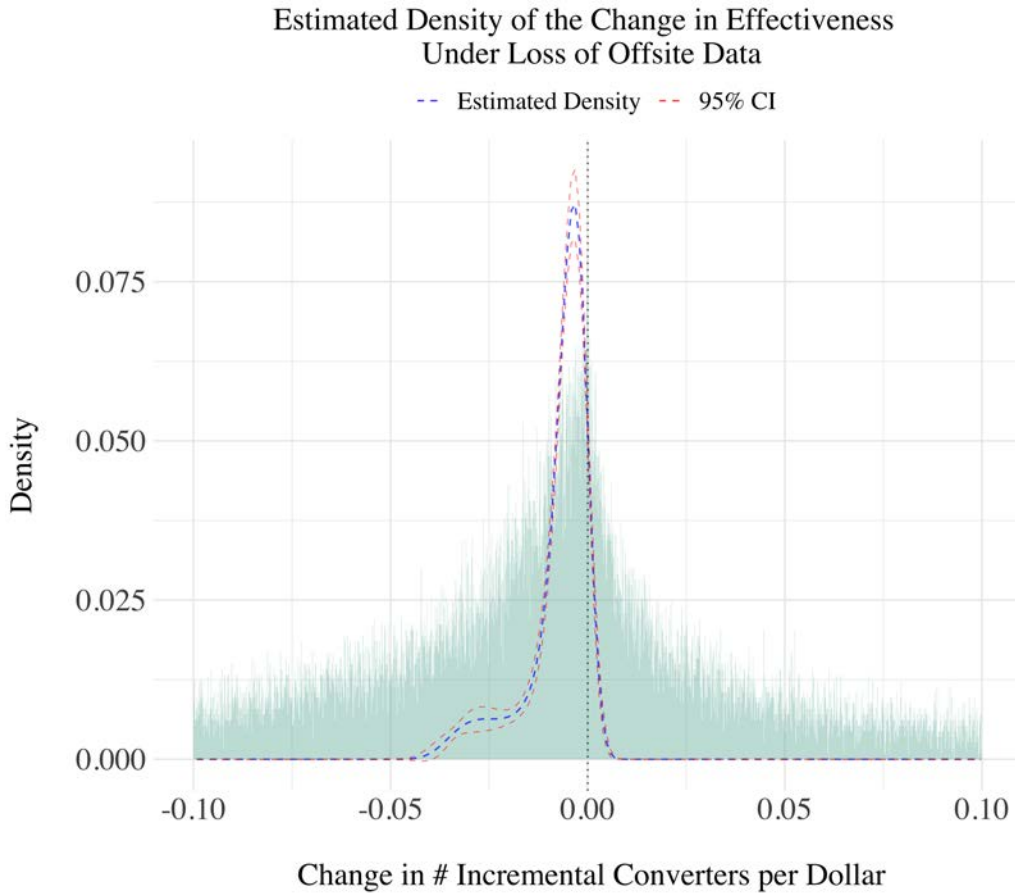
[26] Given the long right tail, the mean results are sensitive to the inclusion/exclusion of the highest percentiles of the empirical treatment effect distribution. Many of those observations are likely spurious, though for reasons unobserved to us (e.g., mis-installed pixels). We prefer the median as it is more robust to such observations.

[27] We note that the overall shape of our estimated distribution is roughly log-normal. Fitting a log-normal to the mean and variance from our estimated distribution yields a median cost per incremental customer of $30.29. Given the support of the log-normal, this cannot capture the roughly 4.5% of our estimated distribution that lies below zero. Back of the envelope, the 45.5th percentile of the log-normal corresponds to a cost per incremental customer of $37.42, which is close to our estimated $38.16. We view this as a rough sanity check on our estimation procedure.

[28] The industry also frequently focuses on a slightly different but related metric – cost per action or acquisition – which often does not capture incrementality. Irvine (2022), for example, reports a median cost per action of $18.68 on Facebook ads from their sample of 256 advertising clients.

[29] Standard errors were calculated from Lemma 2 in Efron (2016). A caveat is that the Fisher information matrix is not guaranteed to be non-negative definite. When that arose in our calculations, we dropped the diagonal term as suggested in Efron (2016).

Figure 8: Estimated distribution of the within-campaign change in the number of incremental converters per dollar.



Notes: The blue line represents the estimated distribution of the within-campaign change in the number of incremental converters per dollar across our sample. The dashed red lines show the 95% confidence intervals calculated from Efron (2016). The empirical histogram of treatment effects is shown in green as a rescaled density with a bin width of 0.001, with the x-axis reduced to (-0.1, 0.1) for display purposes.

Table 8: Quantiles and mean of the estimated distribution of within-campaign change in advertising effectiveness.

	10th	25th	50th	75th	90th	Mean
# Fewer Incremental Converters per $1,000	-20.5	-9.9	-6.2	-2.6	-0.4	-7.5

Notes: The within-campaign change in advertising effectiveness is measured as the reduction in incremental converters per dollar when moving from optimizing for purchases to optimizing for clicks.

Table 8 shows that if campaigns optimized for clicks instead of purchases, the median campaign would generate 6.2 fewer incremental customers for every $1,000 spent. Evaluating this median change relative to the median estimate of baseline ad effectiveness, the cost per incremental converter would increase from $38.16 to $49.93, a 31% increase in costs. To put this in perspective, the median total weekly ad spend in our sample was $1,259. Holding this budget fixed, our observed and counterfactual estimates of cost per incremental customer imply a decrease from 33 to 25 incremental customers per week. This is a large shock to advertising effectiveness: losing the ability to optimize delivery for purchases increases costs substantially in our sample.

Further, the quantiles of the fitted distribution show a lot of mass relatively tightly concentrated below zero; this suggests that it is not just a few firms that would be adversely affected – it is closer to a homogeneous shock across the population. At the same time, we estimate that around 10% of the sample would have generated *more* incremental converters per dollar if they switched to click optimization at the time of our experiment. We believe this relates to the non-incremental nature of Meta's optimization algorithms: purchase-optimized campaigns could be finding users who were already going to purchase the product, whereas optimizing for clicks could cast a wider net and generate more incremental converters. Hence, while there is some heterogeneity across the distribution, on net, the majority of the campaigns in our sample are less effective under our counterfactual.

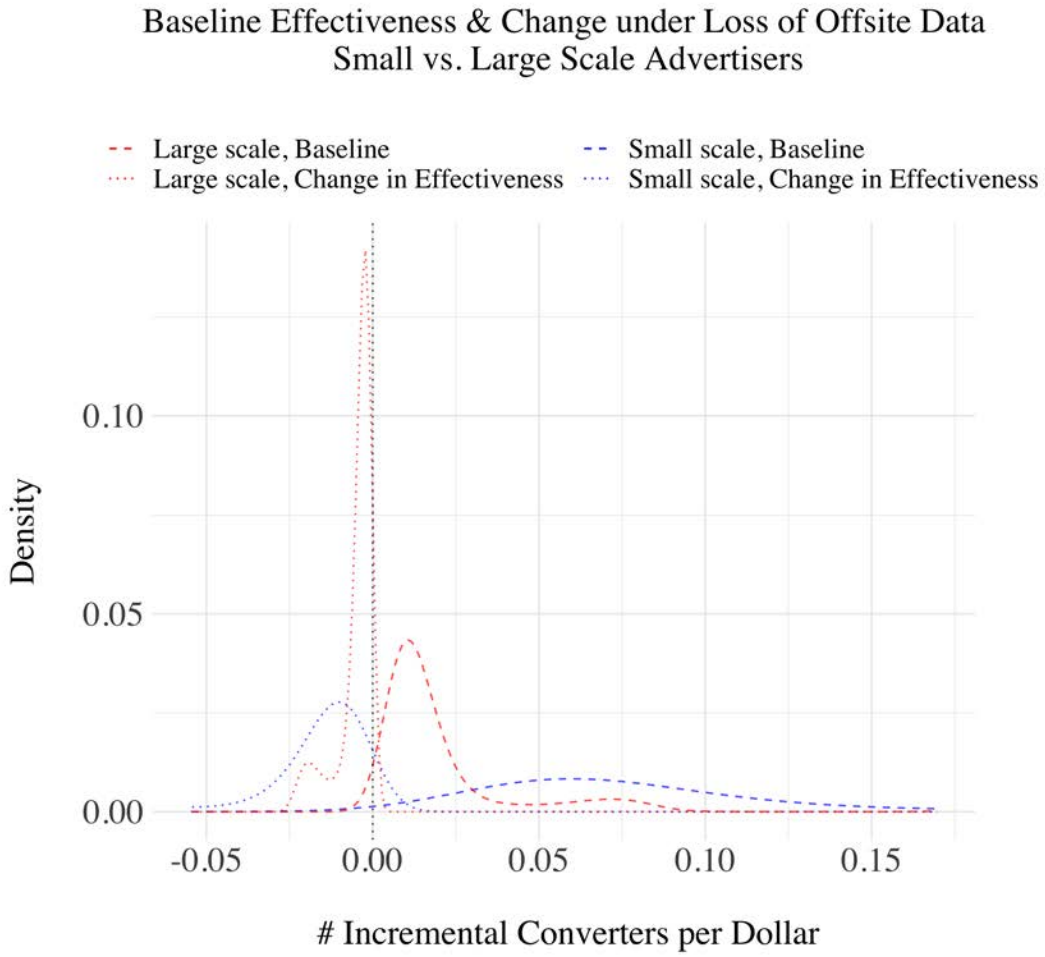
5.2.3 Small vs. Large Scale Advertisers

Small businesses rely heavily on online advertising (Kerrigan and Keating 2019, Herhold 2019). To explore heterogeneity in effects between small and large scale advertisers, we conduct a median split of the sample based on ad spend in the eight weeks prior to the experiment launch.[30] We refer to advertisers below the median as 'small scale' advertisers and those above as 'large scale.'

We re-estimate our main analysis separately for small and large scale advertisers. Figure 9 graphs the resulting four distributions: estimates of the baseline ad effectiveness and within-campaign change for small and large scale advertisers. The blue distributions correspond to small scale advertisers, while the red distributions correspond to large scale advertisers. The dashed lines trace out the densities of the baseline ad effectiveness for these two groups, and the dotted lines correspond to the within-campaign difference. We also include summary statistics of these distributions in Table 9.

[30]This metric is positively correlated with several other metrics around 'firm size': for example, lifetime account spend, number of attributed sales tracked through Meta pixels, and Meta's internal 'small' vs. 'large' classification. The details of Meta's classification scheme are complex and not public. We chose to focus on our selected metric due to a combination of data availability and transparency.

Figure 9: Estimated distribution of effects for small and large scale advertisers.



Notes: The distributions for small scale advertisers are plotted in blue, and the distributions for large scale advertisers are in red. The dashed lines denote the baseline, purchase-optimized ad effectiveness distribution. The dotted lines correspond to the within-campaign change. For interpretability, standard errors are not overlaid, though they follow each distribution closely as in our main results.

Table 9: Quantiles and means for the estimated distributions for small and large scale advertisers.

	10th	25th	50th	75th	90th	Mean
Small Scale Advertisers						
# Incremental Converters per $1,000	27.2	49.4	80.9	153.4	357.4	184.2
Cost per Incremental Converter	$36.74	$20.23	$12.37	$6.52	$2.80	$5.43
# Fewer Incremental Converters per $1,000	-85.1	-57.3	-17.6	-8.1	-1.5	-30.5
Large Scale Advertisers						
# Incremental Converters per $1,000	3.1	7.6	13.5	23.1	66.7	31.1
Cost per Incremental Converter	$318.59	$131.92	$74.06	$43.37	$15.00	$32.14
# Fewer Incremental Converters per $1,000	-15.3	-6.5	-3.8	-2.0	-0.6	-5.0

Notes: The first row in each panel summarizes the effectiveness of purchase-optimized campaigns. The third row in each panel reports the within-campaign change in advertising effectiveness, measured as the change in incremental converters per dollar when moving from optimizing for purchases to optimizing for clicks.

Both distributions are shifted outward for small scale advertisers, indicating that they tend to have more effective ads at baseline and they are hurt more by losing offsite data. Comparing the medians of the within-campaign change in ad effectiveness, we see that the median large scale advertiser loses 3.8 incremental customers per $1,000 spent in our counterfactual whereas the median small scale advertiser loses 17.6.[31,32]

Part of the reason small scale advertisers are hurt more in the counterfactual is that their ads are more effective at baseline: their median cost per incremental converter is $12.37 versus $74.06 for large scale advertisers. While there are many potential reasons for this, we note that this finding is consistent with diminishing returns to advertising. We leave explorations of mechanisms to future research and focus only on the descriptive finding here.

[31] In percentage terms, if one evaluates the median loss at the median ad effectiveness, small scale advertisers face a 28% increase in costs (going from $12.37 = 1/.0809 to $15.82 = 1/(.0809-.0176)) vs. a 39% (going from $74.96 = 1/0.0135 to $103.03 = 1/(0.0135-0.0038)) increase for large scale advertisers. We prefer the absolute comparisons because the median-evaluated-at-the-median can shift from either changes in the numerator or denominator, making clean comparisons hard. In contrast, looking at the distributions of the change in effectiveness, it is clear that the distribution for small scale advertisers is shifted to the left.

[32] An alternative approach would be to incorporate covariates directly into the estimation procedure to estimate a conditional average treatment effect. This can be done by modeling $p_i(X_i \mid \Theta_i) \sim N(\Theta_i + u_i'\gamma\,,\, \hat{\sigma_i}^2)$ for u_i a dummy vector of firm size. We estimated this model for the change in effectiveness and, consistent with our split sample analysis, γ was negative and significant, indicating small scale advertisers were hurt more.

5.2.4 Long Term Effects

Thus far, we have restricted our attention to the count of incremental converters over the course of our week-long experiment. However, not all customers are equally valuable to advertisers: ads delivered with offsite data may generate more short term customers, but that does not necessarily mean they generate more long term customers, who would tend to have higher lifetime valuations for advertisers. Further, studying the long term effects of ads is potentially informative about whether those ads provide benefits to consumers. If an ad induces a user to become a long term customer, that is consistent with a larger benefit on the consumer side than if the ad were to induce either zero or one purchase, since it suggests that the consumer liked the product enough to want to continue buying. If a consumer only buys once after seeing an ad, that could potentially indicate that the consumer was unhappy with their purchase.

We study whether the ads in our experiment generated long-term effects by comparing purchase outcomes across our treatment and holdout groups six months after our initial experiment ran. Since pixels fire regardless of whether users see ads on Meta, we can observe purchasing behavior through pixel fires just as we did during the week of the experiment. Due to data retention issues, we could not look at cumulative purchases over the six month window, so instead we took a snapshot of purchase data during a one week interval six months after our experiment ended.

Focusing on the same campaigns as in our main analysis, we re-ran our estimation procedure, treating this long term purchase data as the outcome variable. Table 10 presents quantiles and means of these distributions. At baseline, the median cost per incremental customer six months in the future is $112.69; under the median change in effectiveness this would rise to $154.77, which is roughly a 37% increase. Hence, there is evidence that offsite data help lower costs to recruit both short and long term customers.

Table 10: Quantiles and means of the estimated distributions for effects measured six months after our experiment.

	10th	25th	50th	75th	90th	Mean
# Incremental Converters per $1,000	0.2	3.8	8.9	20.4	78.6	29.9
Cost per Incremental Converter	$4,532.62	$260.95	$112.69	$49.10	$12.72	$33.43
# Fewer Incremental Converters per $1,000	-4.7	-3.6	-2.4	-1.2	-0.2	-1.9

These results are consistent with, but may not necessarily be indicative of positive consumer surplus. First, our long term outcome is measured six months after random assignment occurred. Because we do not have data from the intervening six month period, we cannot speak to downstream differences across conditions that drive the long term effects. For example, an initial exposure to an ad could lead to subsequent retargeting that generates delayed purchases.[33] Second, repeat purchases could be driven by state-dependence or addiction, which also complicates welfare analyses. These are deeper questions about consumer welfare that we leave for future research.

6 Discussion

We provide experimental estimates of the effectiveness of Meta advertising under both purchase- and click-based optimization across a large-scale, representative set of advertisers. We find considerable heterogeneity in ad effectiveness across advertisers, as well as heterogeneity in the importance of offsite data to these advertisers. These results advance our understanding of digital advertising effectiveness and can be viewed as broadly descriptive in nature. Our findings are also relevant to a number of managerial and policy questions. Many of these are complex questions that our experiment cannot fully address, but we hope our results can deepen the literature's understanding of the issues and inspire further research addressing each in detail.

When it comes to managerial implications, we document a large degree of heterogeneity in the estimated distribution of ad effectiveness at baseline. The cost per incremental customer ranges more than $200 from the 10th to the 90th percentile, and we see a large degree of heterogeneity even within verticals (see Web Appendix F). This fact is consistent with multiple explanations that would provide different implications to advertisers. For example, the variance could be explained by differences in ad quality, which would imply large incentives for firms to invest in ad production. Alternatively, it could be explained by pre-existing differences across products, which would only suggest that different types of products have different optimal advertising spending. As a result, further research that pinpoints the source of the heterogeneous advertising effectiveness we document would be valuable.

Our research also relates to managerial questions surrounding the value of offsite data. A more precise understanding of the value of offsite data can inform investment decisions from managers across several kinds of businesses. For advertisers, our estimates can help

[33] In the case of retargeting, six months is much longer than the retargeting windows used by most advertisers. For example, a Shopify advice column about Facebook retargeting for different actions recommends a max of three months. (Available at https://www.shopify.com/blog/facebook-retargeting.)

inform willingness-to-pay for products across platforms (and data regimes) that use or do not use purchase data in similar ways. For online advertising platforms, our results can inform product strategy. For example, in the wake of Apple's iOS 14.5 change, both Meta and TikTok launched new 'Shops' where consumers can purchase products directly on the social media platforms, and thus advertisers do not have to rely on off-platform data. Evaluating those strategies may be a fruitful area for further research.

On the policy side, future privacy regulations could approach data in different ways. Rather than viewing offsite data as something that should either be 'allowed' or 'restricted' for use in ad targeting (e.g., by blocking the tracking technology entirely), we could alternatively conceptualize this as a policy decision about property rights. The magnitude of and heterogeneity in effects for advertisers we identify, combined with past estimates of how much consumers value their data (e.g., Lin (2022), Athey et al. (2017)), suggest there may be gains from trade for some players. Building out increased markets for an exchange of data between advertisers and users may improve social welfare. Bergemann et al. (2023) provide a thoughtful discussion of the issues around such markets.

When discussing privacy policy, there is also uncertainty about how regulations will impact markets. Our analysis highlights potential directional takeaways for two categories of competitive implications: one stemming from the disparate access to offsite data across advertising platforms and one stemming from the disproportionate effects we observe on small scale advertisers.

With regard to competition between advertising platforms, changes to the offsite data ecosystem may differentially affect platforms and their ability to use such data. For example, with Apple's ATT change, the prompt users had to respond to in order to allow tracking was different for Apple apps than it was for apps made by other companies; this likely led to differential opt in rates for Apple versus other advertising platforms.[34] Changes like this one may generate a competitive advantage for the platforms that retain more tracking data since the ability to target with such data may help attract advertisers. Decreasing the number of platforms with offsite tracking data may reduce options to advertisers and increase the amount advertisers need to spend to advertise on those platforms. While this directional effect appears unambiguous, quantifying the size of such market power changes as a result of decreased data availability is an important area for further research.

With regard to market structure in general product markets, our results that show a large and differential effect on small scale advertisers are highly relevant. To the extent small advertisers rely disproportionately on digital advertising, reduced access to purchase

[34]See, for example, https://assets.publishing.service.gov.uk/media/61b86aeb8fa8f5037778c3b8/Appendix_I_-_Considering_the_impacts_of_Apples_ATT.pdf

optimized ads and likely higher prices on those ads may increase the costs of small firms more than larger competitors. If such changes make it harder for small firms to find customers, it is possible we could see increased concentration in product markets either through exit of existing companies who rely on purchase optimized advertising or through deterred entry based on the anticipated increased cost of finding new customers. This is consistent with a growing literature that documents from both a theoretical and empirical standpoint how privacy regulation can adversely affect small players and entrants (e.g., Campbell et al. (2015), Janssen et al. (2022), Johnson et al. (2023)). Unfortunately, we cannot assess how serious this concern is because our experiment does not evaluate a counterfactual that removes access to offsite purchase data for all advertisers. In particular, it is possible that ad platforms and advertisers could adjust their behavior in such a counterfactual in ways that could mitigate the above hypothesized effects. Estimating general equilibrium effects and computing relative magnitudes of these forces is an important topic for future research.

Finally, we note that the debate over privacy regulation is often framed as a trade-off between consumer privacy and advertising effectiveness for firms, but this need not be the case: companies have started to invest research into privacy-preserving ad delivery and measurement solutions that can comply with privacy regulation while maintaining advertiser-side benefits of the data. Though much of this technology is in its nascent stages, our results can help inform the potential upsides from such R&D and thus motivate commensurate investment by firms or governments. If fully realized, such technology has the potential to both satisfy privacy regulation and help mitigate any effects on advertising effectiveness.

7 Conclusion

We estimate the value of offsite tracking data to advertisers using a large-scale experiment on Meta. Offsite data is believed to be amongst the most important data in digital advertising, and it is used by millions of advertisers around the world. To estimate the first order effects of this data on ad effectiveness, we conducted a large-scale study with more than 70,000 individual experiments that focused on a primary way firms use this data.

We find evidence that the costs of attracting incremental customers are substantially higher when advertisers cannot optimize ad delivery for offsite purchases. The median campaign evaluated at the median loss would experience a 31% increase in costs per incremental customer. We show further evidence that small scale advertisers are disproportionately hurt and that each dollar spent on optimizing delivery for offsite events generates more longer term customers than ads delivered without such data. We discuss the implications of our findings for advertisers, ad platforms, and policymakers.

We hope our research inspires more work on the value of digitized data to advertisers and the associated privacy costs to consumers. The data sharing ecosystem for digital advertising is evolving rapidly and often with scant evidence to inform socially optimal policy or decision-making by the relevant actors. We believe this domain will remain a fruitful area of research for years to come.

Funding and Competing Interests

Wernerfelt and Moakler were employees of Meta when this research was conducted and the latter owns stock in the company. Meta was able to review this publication for proprietary, trade-secret, or non-aggregated information that could potentially identify any individual(s), but did not have the right to restrict publication based on the results or content of the findings.

References

Alcobendas M, Kobayashi S, Shi K, Shum M (2021) The impact of privacy measures on online advertising markets. *Available at SSRN 3782889* .

Angrist JD, Pischke JS (2009) *Mostly harmless econometrics: An empiricist's companion* (Princeton university press).

Athey S, Catalini C, Tucker C (2017) The digital privacy paradox: Small money, small costs, small talk. Technical report, National Bureau of Economic Research.

Athey S, Grabarz K, Luca M, Wernerfelt N (2023) Digital public health interventions at scale: The impact of social media advertising on beliefs and outcomes related to covid vaccines. *Proceedings of the National Academy of Sciences* 120(5):e2208110120.

Bergemann D, Crémer J, Dinielli D, Groh CC, Heidhues P, Schafer M, Schnitzer M, Morton FMS, Seim K, Sullivan M (2023) Market design for personal data. *Yale J. on Reg.* 40:1056.

Campbell J, Goldfarb A, Tucker C (2015) Privacy regulation and market structure. *Journal of Economics & Management Strategy* 24(1):47–73.

Cramer-Flood E (2021) Worldwide digital ad spending 2021. Technical report, eMarketer, URL https://content-na1.emarketer.com/worldwide-digital-ad-spending-2021.

Cramer-Flood E (2023) Worldwide ecommerce forecast 2023. Technical report, eMarketer, URL https://www.insiderintelligence.com/content/worldwide-ecommerce-forecast-2023.

DellaVigna S, Linos E (2022) Rcts to scale: Comprehensive evidence from two nudge units. *Econometrica* 90(1):81–116.

Dubé JP, Misra S (2023) Personalized pricing and consumer welfare. *Journal of Political Economy* 131(1):131–189.

Efron B (2014) *The Bayes deconvolution problem* (Division of Biostatistics, Stanford University Stanford).

Efron B (2016) Empirical bayes deconvolution estimates. *Biometrika* 103(1):1–20.

Goldberg S, Johnson G, Shriver S (2021) Regulating privacy online: An economic evaluation of the GDPR. *Available at SSRN* .

Goldfarb A, Tucker CE (2011) Privacy regulation and online advertising. *Management science* 57(1):57–71.

Gordon BR, Moakler R, Zettelmeyer F (2022) Close enough? a large-scale exploration of non-experimental approaches to advertising measurement. *Marketing Science* .

Gordon BR, Zettelmeyer F, Bhargava N, Chapsky D (2019) A comparison of approaches to advertising measurement: Evidence from big field experiments at facebook. *Marketing Science* 38(2):193–225.

Hackett R, Harty D (2021) Apple's ad changes wiped $142 billion off snap, facebook, and other online ad giants. Technical report, Forbes, URL https://fortune.com/2021/10/22/apple-snap-facebook-earnings-google-twitter-pinterest-ad-tracking/.

Herhold K (2019) The state of small business advertising in 2019. Technical report, The Manifest, URL https://themanifest.com/advertising/small-business-advertising-2019.

Innovid (2021) Reconciling views on identity resolution. Technical report, Innovid, URL https://info.innovid.com/reconciling-views-on-identity-resolution#1.

Irvine M (2022) Facebook ad benchmarks for your industry [data]. URL https://www.wordstream.com/blog/ws/2017/02/28/facebook-advertising-benchmarks.

Janssen R, Kesler R, Kummer ME, Waldfogel J (2022) Gdpr and the lost generation of innovative apps. Technical report, National Bureau of Economic Research.

Jia J, Jin GZ, Wagman L (2018) The short-run effects of gdpr on technology venture investment. Technical report, National Bureau of Economic Research.

Johnson G (2013) The impact of privacy policy on the auction market for online display advertising.

Johnson G, Lewis RA, Nubbemeyer E (2017a) The online display ad effectiveness funnel & carry-over: Lessons from 432 field experiments. *Available at SSRN 2701578* .

Johnson GA, Lewis RA, Nubbemeyer EI (2017b) Ghost ads: Improving the economics of measuring online ad effectiveness. *Journal of Marketing Research* 54(6):867–884.

Johnson GA, Shriver SK, Goldberg SG (2023) Privacy and market concentration: intended and unintended consequences of the gdpr. *Management Science* .

Kalwani MU, Silk AJ (1982) On the reliability and predictive validity of purchase intention measures. *Marketing Science* 1(3):243–286.

Kerrigan K, Keating R (2019) Online advertising delivers big benefits for small businesses. Technical report, Small Business & Entrepreneurship Council, URL https://sbecouncil.org/2019/09/10/online-advertising-delivers-big-benefits-for-small-businesses/.

Kline P, Rose EK, Walters CR (2022) Systemic discrimination among large us employers. *The Quarterly Journal of Economics* 137(4):1963–2036.

Korganbekova M, Zuber C (2023) Balancing user privacy and personalization. *Work in progress* .

Lei X, Chen Y, Sen A (2023) The value of external data for digital platforms: Evidence from a field experiment on search suggestions. *Available at SSRN* .

Lewis RA, Rao JM (2015) The unfavorable economics of measuring the returns to advertising. *Quarterly Journal of Economics* 1941–1973.

Lin T (2022) Valuing intrinsic and instrumental preferences for privacy. *Marketing Science* 41(4):663–681.

Lodish LM, Abraham M, Kalmenson S, Livelsberger J, Lubetkin B, Richardson B, Stevens ME (1995) How tv advertising works: A meta-analysis of 389 real world split cable tv advertising experiments. *Journal of Marketing Research* 32(2):125–139.

Moorman C (2023) The cmo survey. *Durham, NC* .

Narasimhan B, Efron B (2020) deconvolver: A g-modeling program for deconvolution and empirical bayes estimation. *Journal of Statistical Software* 94:1–20.

Novacek G, Black B, Walsh K, Rooney J, Hinchcliffe L (2016) The winner-take-all digital world for cpg. Technical report, Boston Consulting Group, URL https://retailhouse.files.wordpress.com/2016/06/bcg-the-winner-take-all-digital-world-for-cpg-mar-2016_tcm80-206184.pdf.

Rossi PE, McCulloch RE, Allenby GM (1996) The value of purchase history data in target marketing. *Marketing Science* 15(4):321–340.

Sahni NS (2016) Advertising spillovers: Evidence from online field experiments and implications for returns on advertising. *Journal of Marketing Research* 53(4):459–478.

Shapiro BT, Hitsch GJ, Tuchman AE (2021) Tv advertising effectiveness and profitability: Generalizable results from 288 brands. *Econometrica* 89(4):1855–1879.

Sun T, Yuan Z, Li C, Zhang K, Xu J (2023) The value of personal data in internet commerce: A high-stakes field experiment on data regulation policy. *Management Science* .

Tadelis S, Hooton C, Manjeer U, Deisenroth D, Wernerfelt N, Dadson N, Greenbaum L (2023) Learning, sophistication, and the returns to advertising: Implications for differences in firm performance. Technical report, National Bureau of Economic Research.

Walters C (2022) Empirical bayes methods: Theory and application. *NBER Summer Institute Methods Lecture* .

Web Appendix

This Web Appendix has six main sections. First, we outline our main estimation procedure in more detail. Second, we report results from running our analysis on our entire sample, not just the campaigns that hit the recommended weekly minimum conversions in expectation. Third, we describe how we generated our experiment-level treatment effect estimates (and the necessary assumptions on the data). Fourth, we conduct a randomization check on our experimental design. Fifth, we describe an illustrative case study to provide additional intuition behind our experiment. Sixth, we provide results by the three main verticals in our sample (E-commerce, CPG, and Retail).

A Methodology

First, we describe our estimation procedure in more depth. We leverage the approach of Efron (2014, 2016); this section borrows heavily from those sources.

Conditional on our observed distributions of treatment effects and standard errors, our goal is to estimate the true, latent distribution of effects across experiments. Further, we want to do so in as flexible a way as possible, especially given the nonstandard distributions we observe. We start by articulating the general problem at hand and then show how recently developed deconvolution methods can allow us to flexibly estimate our quantities of interest under minimal assumptions.

Suppose from a set of N experiments we observe a set of treatment effects X_1, X_2, ..., X_N. Each X_i is a noisy measure of experiment i's true, unobserved treatment effect Θ_i, and we assume that the Θ_i are distributed according to an unobserved distribution $g(\theta)$. We are interested in making inferences on g based on our realized X_i's. Next, assume that the unobserved distribution of treatment effects are drawn iid from g. That is, $\Theta_i \overset{\text{ind}}{\sim} g(\theta)$ for all $i \in \{1, 2, ..., N\}$. Further, assume that each X_i is drawn independently from Θ_i according to a known distribution p_i: $X_i \overset{\text{ind}}{\sim} p_i(X_i \mid \Theta_i)$. Note that this specifies a hierarchical distribution: first the set of Θ_i are drawn from g, and then conditional on the Θ_i's, we draw our realized X_i's.

Given this set up, if we assume a broad exponential family of models for g, not only can we attain flexible functional forms, but the optimization problem also becomes tractable.[35]

[35] See Kline et al. (2022), Walters (2022) for other recent applications of this methodology. We note Efron (2016) also explores incorporating regularization into the deconvolution estimate; we chose the non-regularized version to

We now lay out that approach and introduce further notation.

Specifically, assume the support of g is a finite discrete set $\mathcal{T} = \{\theta_1, ..., \theta_m\}$. (This assumption is not strictly necessary, but it eases the analysis.) This makes the prior $g(\theta)$ an m-vector $g = (g_1, ..., g_m)$ that specifies the probability g_j on θ_j. We assume:

$$g(\alpha) = exp\{Q\alpha - \phi(\alpha)\} \tag{1}$$

where α is a p-dimensional parameter vector and Q is a known $m \times p$ structure matrix. Denoting by Q_j^T the jth row of Q, we have that the jth component of $g(\alpha)$ is

$$g_j(\alpha) = exp\{Q_j^T\alpha - \phi(\alpha)\} \quad \text{for } j = 1,2,...,m \tag{2}$$

where $\phi(\alpha)$ normalizes $g(\alpha)$ to make it a probability distribution:

$$\phi(\alpha) = \log \sum_{j=1}^{m} exp(Q_j^T\alpha) \tag{3}$$

In our estimation, we follow Narasimhan and Efron (2020) in letting Q be a basis matrix for natural cubic splines over $\mathcal{T}$ with degrees of freedom p. Past applications have analyzed relatively small-scale data compared to ours (e.g., Kline et al. (2022)) and assumed that fixing $p = 5$ granted the model enough underlying flexibility to adequately describe the data. In the event that the true prior is not contained in such a low-dimensional exponential family, the estimated model would contain "definitional bias" in the words of Efron (2016). Empirically, we find that models that assume $p = 5$ fail to capture our skewed distributions well, and so we estimate our models over wide ranges of p and then perform model selection over them. We discuss how we perform model selection later, but empirically our results are robust to several different selection procedures.

The discussion above has focused on the 'higher' level distribution g from which the unobserved Θ_i are drawn. We now turn to the 'lower' level distributions that map the unobserved Θ_i to the observed treatment effects, X_i. Given that context, let

$$p_{ij} = p_i(X_i \,|\, \Theta_i = \theta_j) \tag{4}$$

denote the probability that X_i is realized if $\Theta_i = \theta_j$, and let P_i be the m-vector of probabilities for X_i across all possible values of θ_j: $P_i = (p_{i1}, ..., p_{im})^T$. Importantly, note that the i subscripts on $p_i(X_i \,|\, \Theta_i)$ mean that each experiment can have a different conditional probability distribution over the observed treatment effects. This is important in our setting

avoid any bias in our estimates.

because we not only observe X_i, but we also observe an estimate of the variance, $\hat{\sigma_i}^2$, that varies by i.

For our analysis, we assume that this lower level distribution is normal with known variance from the empirical treatment effect estimate: $p_i(X_i \,|\, \Theta_i) \sim N(\Theta_i\,, \hat{\sigma_i}^2)$. This assumption is an appeal to the Central Limit Theorem; intuitively, this puts no restriction on our higher level distribution of interest, g, but imposes an assumption that our experiments are sufficiently large that our variance estimates have converged and the treatment effects are normally distributed around the unobserved Θ_i.[36]

Given this set up, the marginal probability for X_i becomes:

$$f_i(\alpha) = \sum_{j=1}^{m} p_{ij} g_j(\alpha) = P_i' g(\alpha) \tag{5}$$

and hence the log likelihood function is $l_i(\alpha) = \log P_i' g(\alpha)$. We use maximum likelihood to generate an estimate $\hat{\alpha}$ for α, which then pins down our distribution of interest, g.[37] Note that in this spline-based setup from Efron, performing maximum likelihood over the space of distributions for g ultimately reduces to maximizing over the weights α on the spline bases. This substantially reduces the state space and makes the problem tractable while maintaining a high degree of flexibility.

A.1 Implementation

First, we define a discrete state space $\mathcal{T}$ for the support of $g(\theta)$. Given the wide range of treatment effect estimates, discretizing the entire range from the smallest to largest treatment effect estimate often proved computationally challenging. Hence, we define a grid of bin size 0.001 over the range between the 1st and 99th percentiles of the treatment effect estimates. Given the extreme spread outside the 1st and 99th percentiles of treatment effects, we drop those observations.[38]

[36] For small experiments, this approximation may be less valid. To check robustness along this dimension, we reran our analysis restricting to experiments with more than 50 converters in the experiments themselves; this number has been used as a heuristic for a sample size that is large enough for the Central Limit Theorem to hold (e.g., Angrist and Pischke (2009) discusses some of the evidence around this). This restriction did not change our results substantively, both for our main results and those for the small and large scale advertiser comparison. We also note that this known variance assumption for the lower level distribution is common in meta-analyses, even under different methodologies (e.g., DellaVigna and Linos (2022)).

[37] This approach is sometimes called g-modeling to emphasize that we are interested in the shape of the prior; most empirical Bayes methods are f-modelling in that they are interested in the marginals.

[38] We re-estimated the main results including both the full sample and singleton points for the support of g that coincided with the treatment effects outside of the 1st and 99th percentiles. In these analyses, the median point estimates were unchanged, though the mean increased slightly, as one would expect given the direction of the skew in the underlying distribution. We prefer the results in the main text since we think the extreme observations are likely spurious.

As mentioned previously, we choose Q to be a basis matrix for natural cubic splines with degrees of freedom equal to p. To minimize the risk of rounding errors, we standardize the basis matrix so the columns have mean zero and sum of squares equal to one. We then run models over $p \in \{10, 20, 30, ..., 200\}$. Intuitively, as we vary p, we change the number of knots, thereby adding increased flexibility at the risk of overfitting. We then must choose between models.

A.2 Model selection

Our preferred method of model selection across the possible degrees of freedom relies on cross-validation. However, we note that under several different model selection procedures (e.g., Bayesian Information Criterion or Akaike Information Criterion) our results do not change substantively. We now describe how we performed our cross-validation procedure.

For a given degrees of freedom, p, we partition the set of experiments into k subsets. We use $k-1$ folds to estimate a prior g and then evaluate the likelihood of observing the treatment effects in the k^{th} fold under that prior. We repeat this process k times for each degree of freedom, so each fold is used $k-1$ times for training and once for evaluation. For each p, this process generates an average value of out of sample performance across folds; we can then select the p that performs best. For each meta-analysis, we then re-estimate the model on the full sample with the selected degrees of freedom. Our main results (Table 7 and Table 8) are from models selected with ten-fold cross-validation; due to computational constraints our remaining results are based on three-fold cross-validation. We note that this means for every distribution we estimate in the main text, we fit 61 different models (20 values of p $\times$ three-fold cross-validation for each p plus 1 final model estimation from selected p value).

A.3 Defining advertisers and campaigns

For the purposes of our experiments, we treat each pixel as a single advertiser, and we treat all ads that are optimized against that pixel as a single campaign. Each experiment is defined at the campaign level, meaning the final number of advertisers, experiments, campaigns, and pixels are equal.

This definition is usually straightforward to implement. However, a wrinkle is that multiple ad accounts can run ads optimizing against the same pixel. This is often done when, for example, a company runs ads from different departments or works with external agencies. In framing the total number of 'advertisers' in our study, we thus did not use the number of advertising accounts, but rather the number of pixels as that seemed a better, if more

conservative, estimate.

Some of our analyses use advertiser-level characteristics. Because there are instances where multiple accounts were used for the same pixel but differed in one of the characteristics (e.g., country), we map account characteristics to advertisers by taking the modal demographic across the accounts within each experiment, weighted by delivered ad impressions. For example, if an experiment involved three accounts and the plurality of impressions came from accounts in the US, we would label the experiment as from the US.

In practice, 85% of experiments have one account, meaning the aforementioned wrinkles affected a relatively small fraction of our sample. Further, rerunning our analyses restricting to accounts that have this one-to-one mapping with experiments yields very similar results.

B Results from the Full Sample

In the main text we restrict our analyses to campaigns that hit the recommended number of conversions per week in expectation; here, we re-run our results on our entire sample. This consists of 150,757 experiments versus the 70,909 in the main text.[39] Figure A1 displays both the distribution of baseline ad effectiveness and the distribution of the change in effectiveness when optimizing for clicks instead of purchases.

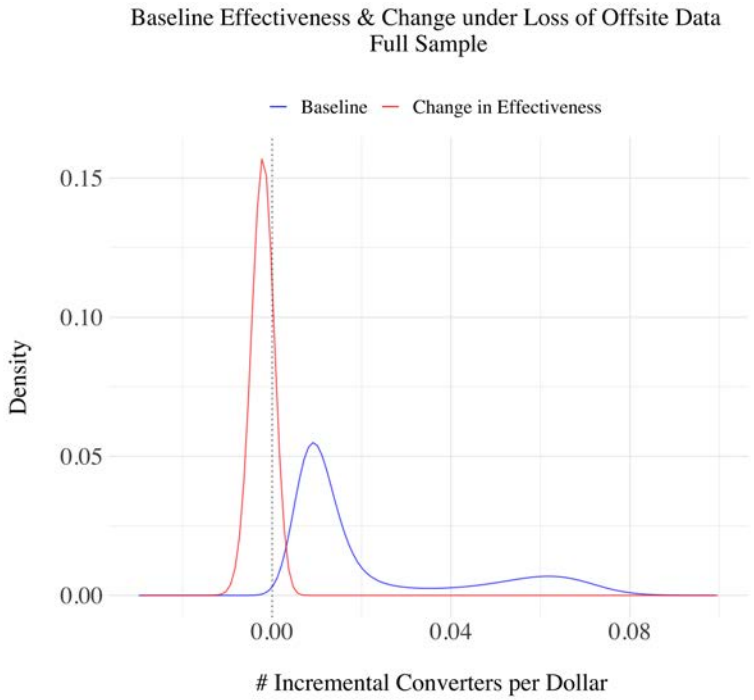
Table A1: Quantiles and means from the estimated baseline distribution of advertising effectiveness.

	10th	25th	50th	75th	90th	Mean
# Incremental Converters per $1,000	5.1	8.1	13.5	53.1	206.6	48.4
Cost per Incremental Converter	$195.03	$123.04	$74.13	$18.82	$4.84	$20.65
# Fewer Incremental Converters per $1,000	-6.0	-4.3	-2.6	-0.9	-0.3	-2.3

Intuitively, given that the optimization algorithm does not perform as well for the additional campaigns that we now retain in the sample, we would expect the cost per incremental converter to increase compared to the main text. We confirm that this is the case – if we recompute the median change at the median effectiveness, we get that cost per incremental

[39] Of the initial 187,922, we drop experiments outside the 1st and 99th percentiles of the treatment effects. We also drop 33,416 (18%) experiments that have no recorded purchases in either holdout or exposed; these treatment effects have zero variance, which is incompatible with our maximum likelihood estimation procedure. We note if you assume all these experiments generate no incremental converters per dollar and are not affected by losing offsite data, and recompute the percentiles of each distribution, the median cost per incremental converter is $105.00 that under the median change in effectiveness would increase to $124.14, an 18% increase.

Figure A1: Estimated distributions from our full sample of the baseline number of incremental converters per dollar and the within-campaign change.



Notes: The blue line traces the baseline ad effectiveness distribution. The red line corresponds to the distribution of the within-campaign change in ad effectiveness. The dashed black lines represent the 95% confidence intervals.

converter increases from $74.13 to $92.23, a 24% increase. This is less than the 31% we estimated for our main sample, but still a substantial increase. Notably, the baseline distribution appears bimodal; we suspect this is due to an integer-related point because many of the advertisers who did not meet the minimum threshold were spending small amounts, and so their experiments had fewer users in them. With fewer people in the denominator, across experiments the cost per incremental converter is not uniformly distributed across the range but rather clusters at certain fractions (e.g., 1/30, 1/20).

C Generation of Experiment-level estimates

As input into our meta-analysis, we first need to generate estimates of the treatment effects and standard errors for each individual experiment. In this section, we walk through how we went from the individual-level data to experiment-level results.

First, without loss of generality, we focus on the baseline case (the derivation for the click optimized arm is identical). In the baseline case, define:

$$
\begin{aligned}
N_H &:= \text{the number of users in the Holdout group} \\
N_E &:= \text{the number of users in the Exposed group} \\
N_H^C &:= \text{the number of converters in the Holdout group} \\
N_E^C &:= \text{the number of converters in the Exposed group} \\
d &:= \text{the ad spend in dollars in the Exposed group}
\end{aligned}
$$

Given those definitions, we constructed our main outcome variable, the number of incremental converters per dollar, as:

$$\tilde{N}_I := \frac{1}{d}\left(N_E^C - N_H^C\left(\frac{N_E}{N_H}\right)\right)$$

This expression takes the number of converters in each of the Holdout and Exposed groups, scales them by the number of users in each, and then divides by the spend.

In terms of variance, we first define the variance for the number of converters in the Holdout group (defined analogously for the number in the Exposed group):

$$\sigma_H^2 := N_H\left(\frac{N_H^C}{N_H} - \left(\frac{N_H^C}{N_H}\right)^2\right)$$

The above is derived from the formula for the variance of a binomial distribution, $np(1-p)$, with N_H independent trials and probability of success N_H^C/N_H. The key assumption is that each user represents an independent trial.

Letting σ_E^2 denote the corresponding quantity for the Exposed group, we get that the variance of the number of incremental converters per dollar is:

$$\tilde{\sigma}^2 := \frac{1}{d^2}\left(\sigma_E^2 + \sigma_H^2\left(\frac{N_E}{N_H}\right)^2\right)$$

The above formulas define our treatment effect and standard error for our baseline case. We now turn to those for the change in effectiveness. For this, we need to introduce subscripts referring to the purchase and click optimized conditions.

Let $\tilde{N}_{purchase}$ denote the number of incremental converters per dollar of the purchase optimized arm and $\tilde{N}_{click}$ denote that for the click arm. We then define our main outcome variable of interest as the change in the number of incremental converters per dollar across arms as:

$$\Delta\tilde{N} := \tilde{N}_{click} - \tilde{N}_{purchase}$$

For the variance of this estimate, let $\tilde{\sigma}^2_{purchase}$ denote the variance of the number of incremental converters per dollar in the purchase arm and $\tilde{\sigma}^2_{click}$ denote that for the click arm. We then define the variance of the difference as:

$$\tilde{\sigma}^2_{\Delta} := \tilde{\sigma}^2_{purchase} + \tilde{\sigma}^2_{click} - 2\text{Cov}(\tilde{N}_{click}, \tilde{N}_{purchase})$$

Under the assumption each user is an independent trial, the covariance term drops out, leaving us with only the first two terms.

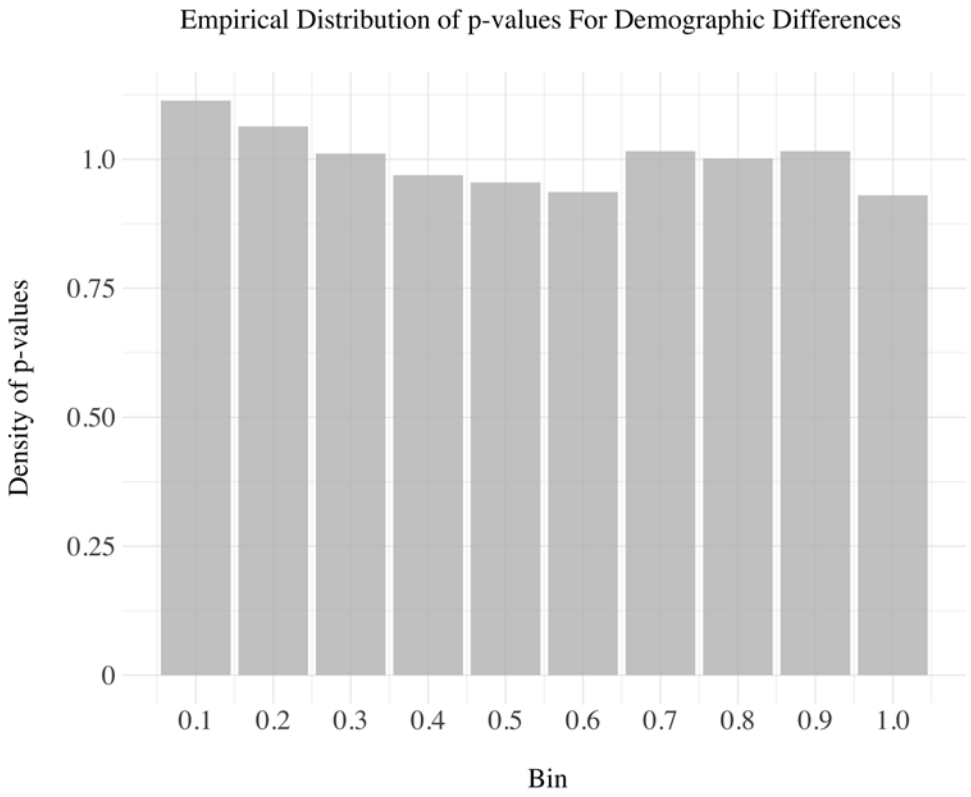
D Randomization Check

As mentioned in the main text, our randomization was done using the same infrastructure that underlies Meta's advertiser-facing lift product. This technology has previously been described in several papers, each of which show that the experiments that were conducted passed randomization checks (Gordon et al. 2022, Athey et al. 2023, Gordon et al. 2019). In this appendix, we report randomization checks on our data as well.

To conduct randomization checks, we collected data on user demographics for a random sample of experiments in our study. We selected demographics that have good coverage in the data and that have standardized values to facilitate analysis: age, gender, account age, number of Facebook friends, iOS user indicator, and web and mobile activity on Facebook in the past month. We restrict to studies with more than 50 users in the control to avoid small sample sizes that could skew our randomization tests. We conducted this check about a year and a half after the experiment was run which means these demographics come from a later point in time. However, while some users may have dropped off or values of some of these demographics may have shifted, any changes seen in the test group we would expect to be mirrored in the control group.

For each of the randomly selected experiments, we conducted t-tests across the test and control groups for each demographic feature. We expect the resulting distribution of p-values will be uniform on [0,1] under correct randomization. Figure A2 plots the density of p-values across 10 evenly spaced bins over [0,1]. Across bins, 6% of p-values are below 0.05, 51% are below 0.50, and 75% are below 0.75, consistent with correct randomization. A chi-squared test comparing our distribution of p-values to what we would expect if they were uniformly distributed leads us to fail to reject the null hypothesis that there are no differences between the distributions, $\chi^2(9, N = 6,918) = 9.94, p = .355$.

Figure A2: Empirical distribution of p-values from randomization checks.



Notes: For each randomly selected experiment, we conduct a t-test for the equality of means of the demographics of users in the treatment and control groups. The distribution reports the resulting p-values, pooled across demographic variables and experiments. Under the null of correct randomization, we would expect to see a uniform distribution of p-values.

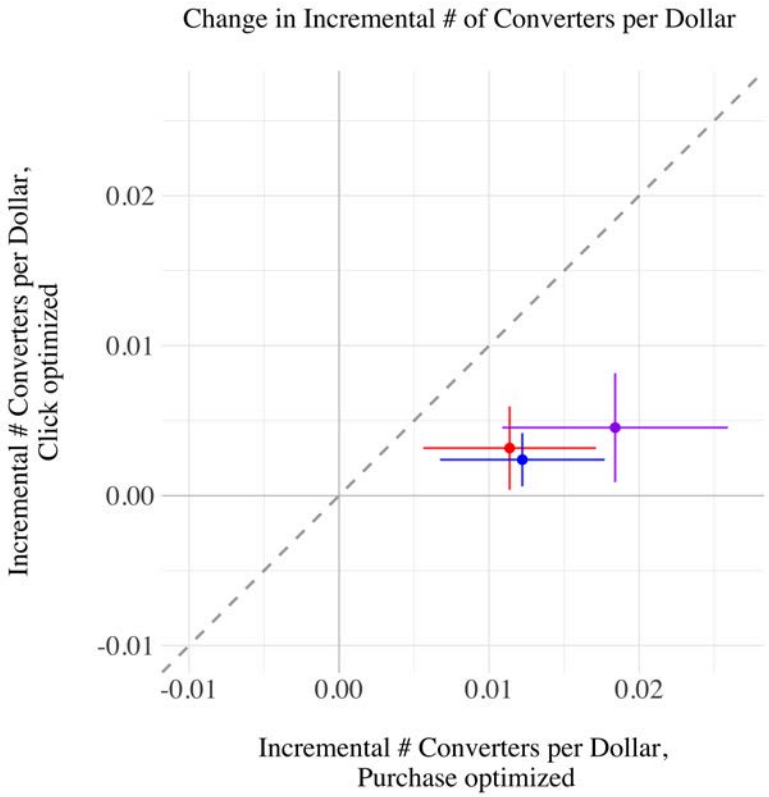
Finally, we end with a caveat to our randomization check. Namely, given these demographics were collected after our experiment was run, if our treatment induced differential attrition on the platform or changed any of our demographics, there could be true imbalances at the time of the experiment that this check fails to detect. Given the small effect sizes we observe in our sample and the fact that several past papers have conducted independent validation of Meta's lift architecture, we are not overly concerned about this caveat but flag it nonetheless.

E Illustrative Case Study and Visualization

To help build intuition behind our experimental design, we first visualize the results from a few select campaigns, and then we visualize the results from the full sample that we use in our main analysis.

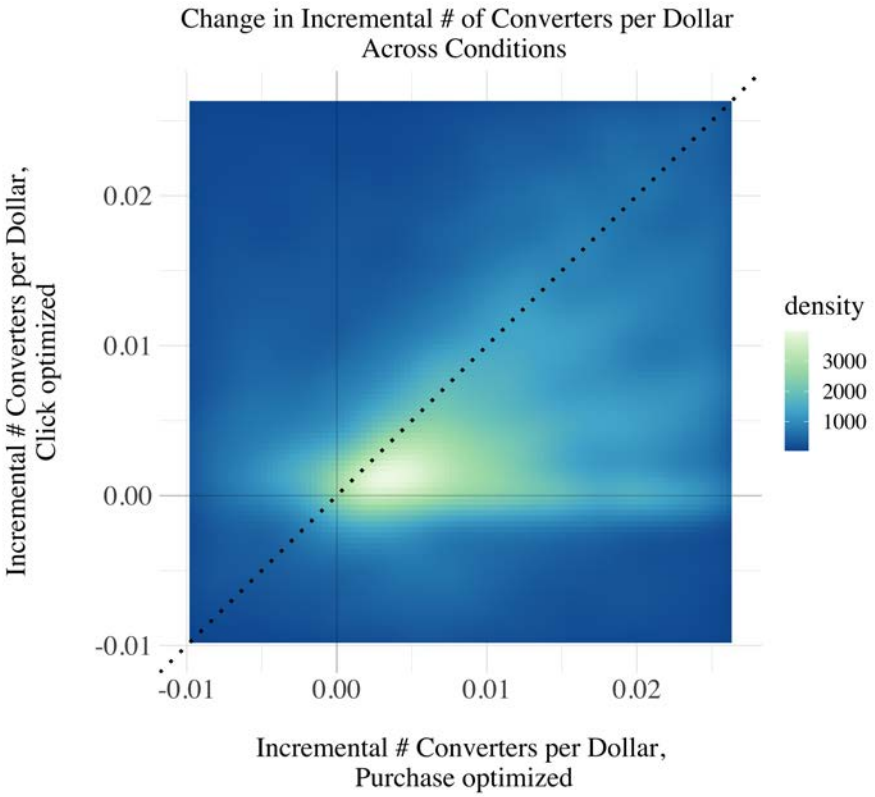
First, in Figure A3, we focus on three example campaigns, and we plot their treatment effects and 95% confidence intervals at baseline (x-axis) and in the counterfactual (y-axis). The red dot represents an apparel company, the blue dot represents a beauty company, and the purple dot represents a jewelry company. The fact that these point estimates lie to the right of the y-axis indicates that, at baseline, each of the campaigns generated a positive number of incremental converters per dollar when the campaigns were optimizing for purchases. Similarly, the fact that they lie below the 45 degree line indicates that they are estimated to earn fewer incremental converters per dollar when they optimize for clicks instead of purchases. For example, we estimate that the jewelry company's (purple dot) cost per incremental converter increased from \$54 to \$154.

Figure A3: Example of treatment effects for three hand-picked advertisers.



In Figure A4, we expand this analysis, showing a heatmap that includes all the treatment effect estimates in our main sample. (This heatmap is similar in spirit to L'Abbé plots that are often used in meta-analyses.) We can clearly see the density to the right of the y-axis and below the 45 degree line: the mass of our estimated treatment effects across all our experiments suggests both a positive baseline effect and a negative effect of losing pixel data.

Figure A4: Heatmap of all 70,909 treatment effects in our main sample.



F Results by Firm Vertical

Here we explore how the results differ by the three largest verticals in our sample: E-commerce, CPG, and Retail. These numbers help provide a sense of heterogeneity across advertisers and may also be useful for practitioners.

Among these three verticals, CPG advertisers have the highest median cost per incremental customer at baseline. As this is a purely descriptive exercise, we cannot say whether that is due to differences in products sold, advertising quality, share of offline sales[40], or other unobservables that differ across verticals. Separately, we also note the change in effectiveness (in an absolute sense) is similar across the verticals.

Table A2: Quantiles and means of estimated distributions by vertical.

	10th	25th	50th	75th	90th	Mean
E-commerce						
# Incremental Converters per $1,000	5.8	12.0	21.5	68.3	94.3	56.4
Cost per Incremental Converter	$172.50	$83.05	$46.52	$14.64	$10.60	$17.74
# Fewer Incremental Converters per $1,000	-15.0	-10.0	-6.4	-3.3	-1.0	-7.7
Consumer Packaged Goods						
# Incremental Converters per $1,000	4.5	9.5	16.7	39.7	95.7	43.0
Cost per Incremental Converter	$222.72	$105.52	$59.89	$25.19	$10.45	$23.25
# Fewer Incremental Converters per $1,000	-12.2	-8.9	-5.9	-3.2	-0.9	-6.5
Retail						
# Incremental Converters per $1,000	6.7	16.9	33.4	132.2	213.2	115.0
Cost per Incremental Converter	$149.16	$59.21	$29.96	$7.57	$4.69	$8.70
# Fewer Incremental Converters per $1,000	-11.6	-8.4	-5.4	-2.5	-2.5	-0.1

[40] We note that a large majority of CPG and Retail sales still occur offline (e.g., Boston Consulting Group estimates that 95% of US CPG sales occur offline (Novacek et al. 2016); eMarketer similarly estimates 81% of global retail sales occur offline Cramer-Flood (2023)). Our estimates of ad effectiveness are biased downward to the extent that the online ads in our study drove offline sales, suggesting there may be more scope for bias in CPG and Retail. At the same time, we note that the vertical with nearly all of its sales online, E-commerce, experiences the largest magnitude change in effectiveness, albeit by a small margin. We leave a deeper exploration of these points to future work.